Signal Processing for Music Analysis

Meinard Müller, *Member, IEEE*, Daniel P. W. Ellis, *Senior Member, IEEE*, Anssi Klapuri, *Member, IEEE*, and Gaël Richard, *Senior Member, IEEE*

***Abstract*—Music signal processing may appear to be the junior relation of the large and mature field of speech signal processing, not least because many techniques and representations originally developed for speech have been applied to music, often with good results. However, music signals possess specific acoustic and structural characteristics that distinguish them from spoken language or other nonmusical signals. This paper provides an overview of some signal analysis techniques that specifically address musical dimensions such as melody, harmony, rhythm, and timbre. We will examine how particular characteristics of music signals impact and determine these techniques, and we highlight a number of novel music analysis and retrieval tasks that such processing makes possible. Our goal is to demonstrate that, to be successful, music audio signal processing techniques must be informed by a deep and thorough insight into the nature of music itself.**

***Index Terms*—Beat, digital signal processing, harmony, melody, music analysis, music information retrieval, music signals, pitch, rhythm, source separation, timbre, voice separation.**

I. Introduction

MUSIC is a ubiquitous and vital part of the lives of billions of people worldwide. Musical creations and performances are among the most complex and intricate of our cultural artifacts, and the emotional power of music can touch us in surprising and profound ways. Music spans an enormous range of forms and styles, from simple, unaccompanied folk songs, to orchestras and other large ensembles, to a minutely constructed piece of electronic music resulting from months of work in the studio.

The revolution in music distribution and storage brought about by personal digital technology has simultaneously fueled tremendous interest in and attention to the ways that information technology can be applied to this kind of content. From browsing personal collections, to discovering new artists, to managing and protecting the rights of music creators, computers are now deeply involved in almost every aspect of music consumption, which is not even to mention their vital role in much of today's music production.

This paper concerns the application of signal processing techniques to music signals, in particular to the problems of analyzing an existing music signal (such as piece in a collection) to extract a wide variety of information and descriptions that may be important for different kinds of applications. We argue that there is a distinct body of techniques and representations that are molded by the particular properties of music audio—such as the pre-eminence of distinct fundamental periodicities (pitches), the preponderance of overlapping sound sources in musical ensembles (polyphony), the variety of source characteristics (timbres), and the regular hierarchy of temporal structures (beats). These tools are more or less unlike those encountered in other areas of signal processing, even closely related fields such as speech signal processing. In any application, the more closely the processing can reflect and exploit the particular properties of the signals at hand, the more successful it will be. Musical signals, despite their enormous diversity, do exhibit a number of key properties that give rise to the techniques of music signal processing, as we shall see.

The application of signal processing to music signals is hardly new, of course. It could be argued to be the basis of the theremin, a 1920s instrument in which an oscillator is controlled by the capacitance of the player's hands near its antennae. The development of modern signal processing in the 1940s and 1950s led directly the first wave of electronic music, in which composers such as Karlheinz Stockhausen created music using signal generators, ring modulators, etc., taken straight from electronics labs. Following the advent of general-purpose digital computers in the 1960s and 1970s, it was not long before they were used to synthesize music by pioneers like Max Matthews and John Pierce. Experimental music has remained a steady source of innovative applications of signal processing, and has spawned a significant body of sophisticated techniques for synthesizing and modifying sounds [1], [2].

As opposed to synthesis, which takes a compact, abstract description such as a musical score and creates a corresponding signal, our focus is analysis—for instance, the inverse problem of recovering a score-level description given only the audio. It turns out that this problem is very computationally demanding, and although efforts at automatic transcription, for example, date back to the mid 1970s [3], the vastly improved computational resources of recent years, along with the demands and opportunities presented by massive online music collections, have led to a recent explosion in this research. The first International Symposium on Music Information Retrieval[1] was held in 2000; this annual meeting is now a thriving interdisciplinary community with over 100 papers presented at the 2010 conference in Utrecht.

Manuscript received September 27, 2010; accepted December 03, 2010. Date of publication February 04, 2011; date of current version nulldate. The work of M. Müller was supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI). The associate editor coordinating the review of this manuscript and approving it for publication was XXXX XXXXXX.

M. Müller is with the Saarland University, 66123 Saarbrücken, Germany, and also with the Max-Planck Institut für Informatik, 66123 Saarbrücken, Germany (e-mail: meinard@mpi-inf.mpg.de).

D. P. W. Ellis is with Electrical Engineering Department, Columbia University, New York, NY 10027 USA (e-mail: dpwe@ee.columbia.edu).

A. Klapuri is with Queen Mary University of London, London E1 4NS, U.K. (e-mail: anssi.klapuri@elec.qmul.ac.uk).

G. Richard is with the Institut TELECOM, Télécom ParisTech, CNRS-LTCI, F-75634 Paris Cedex 13, France (e-mail: gael.richard@telecom-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2011.2112333

[1][Online]. Available: http://www.ismir.net/.

In the rest of the paper, we will present several of the distinctive aspects of the musical signal, and the most significant approaches that have been developed for its analysis. While many techniques are initially borrowed from speech processing or other areas of signal processing, the unique properties and stringent demands of music signals have dictated that simple repurposing is not enough, leading to some inspired and elegant solutions. We have organized the material along particular musical dimensions: In Section II, we discuss the nature of pitch and harmony in music, and present time–frequency representations used in their analysis. Then, in Section III, we address the musical aspects of note onsets, beat, tempo, and rhythm. In Section IV, we discuss models representing the timbre and instrumentation of music signals and introduce various methods for recognizing musical instruments on audio recordings. Finally, in Section V, we show how acoustic and musical characteristics can be utilized to separate musical voices, such as the melody and bass line, from polyphonic music. We conclude in Section VI with a discussion of open problems and future directions.

II. Pitch and Harmony

Pitch is a ubiquitous feature of music. Although a strict definition of music is problematic, the existence of sequences of sounds with well-defined fundamental periods—i.e., individual notes with distinct pitches—is a very common feature. In this section, we discuss the objective properties of musical pitch and its use to create musical harmony, then go on to present some of the more common time–frequency representations used in music signal analysis. A number of music applications based on these representations are described.

A. Musical Pitch

Most musical instruments—including string-based instruments such as guitars, violins, and pianos, as well as instruments based on vibrating air columns such as flutes, clarinets, and trumpets—are explicitly constructed to allow performers to produce sounds with easily controlled, locally stable fundamental periods. Such a signal is well described as a harmonic series of sinusoids at multiples of a fundamental frequency, and results in the percept of a musical note (a single perceived event) at a clearly defined pitch in the mind of the listener. With the exception of unpitched instruments like drums, and a few inharmonic instruments such as bells, the periodicity of individual musical notes is rarely ambiguous, and thus equating the perceived pitch with fundamental frequency is common.

Music exists for the pleasure of human listeners, and thus its features reflect specific aspects of human auditory perception. In particular, humans perceive two signals whose fundamental frequencies fall in a ratio 2:1 (an octave) as highly similar [4] (sometimes known as "octave equivalence"). A sequence of notes—a melody—performed at pitches exactly one octave displaced from an original will be perceived as largely musically equivalent. We note that the sinusoidal harmonics of a fundamental at f_0 at frequencies $f_0, 2f_0, 3f_0, 4f_0, \ldots$ are a proper superset of the harmonics of a note with fundamental $2f_0$ (i.e., $2f_0, 4f_0, 6f_0, \ldots$), and this is presumably the basis of the perceived similarity. Other pairs of notes with frequencies in simple ratios, such as f_0 and $3f_0/2$ will also share many harmonics, and are also perceived as similar—although not as close as the octave. Fig. 1 shows the waveforms and spectrograms of middle C (with fundamental frequency 262 Hz) played on a piano and a violin. Zoomed-in views above the waveforms show the relatively stationary waveform with a 3.8-ms period in both cases. The spectrograms (calculated with a 46-ms window) show the harmonic series at integer multiples of the fundamental. Obvious differences between piano and violin sound include the decaying energy within the piano note, and the slight frequency modulation ("vibrato") on the violin.

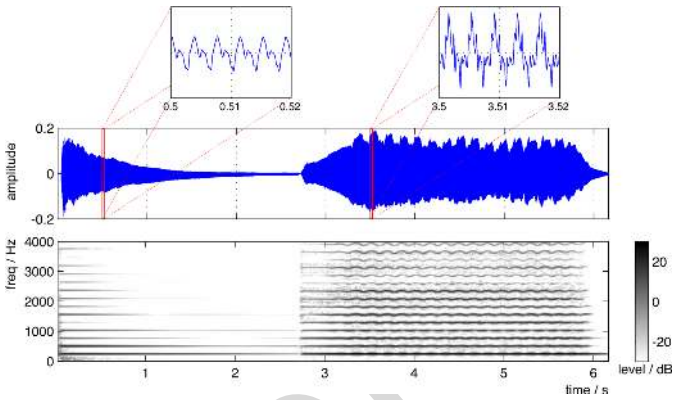

Fig. 1. Middle C (262 Hz) played on a piano and a violin. The top pane shows the waveform, with the spectrogram below. Zoomed-in regions shown above the waveform reveal the 3.8-ms fundamental period of both notes.

Although different cultures have developed different musical conventions, a common feature is the musical "scale," a set of discrete pitches that repeats every octave, from which melodies are constructed. For example, contemporary western music is based on the "equal tempered" scale, which, by a happy mathematical coincidence, allows the octave to be divided into twelve equal steps on a logarithmic axis while still (almost) preserving intervals corresponding to the most pleasant note combinations. The equal division makes each frequency $2^{1/12} \approx 1.06\times$ larger than its predecessor, an interval known as a semitone. The coincidence is that it is even possible to divide the octave uniformly into such a small number of steps, and still have these steps give close, if not exact, matches to the simple integer ratios that result in consonant harmonies, e.g., $\left(2^{1/12}\right)^7 = 1.498 \approx 3/2$, and $\left(2^{1/12}\right)^5 = 1.335 \approx 4/3$. The western major scale spans the octave using seven of the twelve steps—the "white notes" on a piano, denoted by C, D, E, F, G, A, B. The spacing between successive notes is two semitones, except for E/F and B/C which are only one semitone apart. The "black notes" in between are named in reference to the note immediately below (e.g., C♯), or above (D♭), depending on musicological conventions. The octave degree denoted by these symbols is sometimes known as the pitch's *chroma*, and a particular pitch can be specified by the concatenation of a chroma and an octave number (where each numbered octave spans C to B). The lowest note on a piano is A0 (27.5 Hz), the highest note is C8 (4186 Hz), and middle C (262 Hz) is C4.

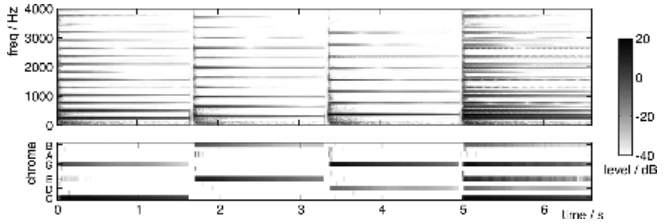

Fig. 2. Middle C, followed by the E and G above, then all three notes together—a C Major triad—played on a piano. Top pane shows the spectrogram; bottom pane shows the chroma representation.

B. Harmony

While sequences of pitches create melodies—the "tune" of a music, and the only part reproducible by a monophonic instrument such as the voice—another essential aspect of much music is harmony, the simultaneous presentation of notes at different pitches. Different combinations of notes result in different musical colors or "chords," which remain recognizable regardless of the instrument used to play them. Consonant harmonies (those that sound "pleasant") tend to involve pitches with simple frequency ratios, indicating many shared harmonics. Fig. 2 shows middle C (262 Hz), E (330 Hz), and G (392 Hz) played on a piano; these three notes together form a C Major triad, a common harmonic unit in western music. The figure shows both the spectrogram and the chroma representation, described in Section II-E below. The ubiquity of simultaneous pitches, with coincident or near-coincident harmonics, is a major challenge in the automatic analysis of music audio: note that the chord in Fig. 2 is an unusually easy case to visualize thanks to its simplicity and long duration, and the absence of vibrato in piano notes.

C. Time–Frequency Representations

Some music audio applications, such as transcribing performances, call for explicit detection of the fundamental frequencies present in the signal, commonly known as pitch tracking. Unfortunately, the presence of multiple, simultaneous notes in polyphonic music renders accurate pitch tracking very difficult, as discussed further in Section V. However, there are many other applications, including chord recognition and music matching, that do not require explicit detection of pitches, and for these tasks several representations of the pitch and harmonic information—the "tonal content" of the audio—commonly appear. Here, we introduce and define these basic descriptions.

As in other audio-related applications, the most popular tool for describing the time-varying energy across different frequency bands is the short-time Fourier transform (STFT), which, when visualized as its magnitude, is known as the spectrogram (as in Figs. 1 and 2). Formally, let x be a discrete-time signal obtained by uniform sampling of a waveform at a sampling rate of F_s Hz. Using an N-point tapered window w (e.g., Hamming $w(n) = 0.54 - 0.46\cos(2\pi n/N)$ for $n \in [0:N-1] := \{0,1,\ldots N-1\}$) and an overlap of half a window length, we obtain the STFT

$$X(t,k) = \sum_{n=0}^{N-1} w(n)x(n+t\cdot N/2)\exp\{-j2\pi kn/N\} \quad (1)$$

with $t \in [0:T-1]$ and $k \in [0:K]$. Here, T determines the number of frames, $K = N/2$ is the index of the last unique frequency value, and thus $X(t,k)$ corresponds to the window beginning at time $t \cdot N/(2F_s)$ in seconds and frequency

$$f_{\text{coeff}}(k) = \frac{k}{N}\cdot F_s \quad (2)$$

in Hertz (Hz). Typical values of $F_s = 44\,100$ and $N = 4096$ give a window length of 92.8 ms, a time resolution of 46.4 ms, and frequency resolution of 10.8 Hz.

$X(t,k)$ is complex-valued, with the phase depending on the precise alignment of each short-time analysis window. Often it is only the magnitude $|X(t,k)|$ that is used. In Figs. 1 and 2, we see that this spectrogram representation carries a great deal of information about the tonal content of music audio, even in Fig. 2's case of multiple, overlapping notes. However, close inspection of the individual harmonics at the right-hand side of that figure—at 780 and 1300 Hz, for example—reveals amplitude modulations resulting from phase interactions of close harmonics, something that cannot be exactly modeled in a magnitude-only representation.

D. Log-Frequency Spectrogram

As mentioned above, our perception of music defines a logarithmic frequency scale, with each doubling in frequency (an octave) corresponding to an equal musical interval. This motivates the use of time–frequency representations with a similar logarithmic frequency axis, which in fact correspond more closely to representation in the ear [5]. (Because the bandwidth of each bin varies in proportion to its center frequency, these representations are also known as "constant-Q transforms," since each filter's effective center frequency-to-bandwidth ratio—its Q—is the same.) With, for instance, 12 frequency bins per octave, the result is a representation with one bin per semitone of the equal-tempered scale.

A simple way to achieve this is as a mapping applied to an STFT representation. Each bin in the log-frequency spectrogram is formed as a linear weighting of corresponding frequency bins from the original spectrogram. For a log-frequency axis with K_L bins, this calculation can be expressed in matrix notation as $\mathbf{Y} = \mathbf{M}\mathbf{X}$, where $\mathbf{Y}$ is the log-frequency spectrogram with K_L rows and T columns, $\mathbf{X}$ is the original STFT magnitude array $|X(t,k)|$ (with t indexing columns and k indexing rows). $\mathbf{M}$ is a weighting matrix consisting of K_L rows, each of $K+1$ columns, that give the weight of STFT bin $|X(\cdot,k)|$ contributing to log-frequency bin $Y(\cdot,\ell)$. For instance, using a Gaussian window

$$M(\ell,k) = \exp\left\{-\frac{1}{2B^2}\left(\log_2\frac{f_{\text{coeff}}(k)}{f_{\min}} - \frac{\ell}{N_O}\right)^2\right\} \quad (3)$$

where B defines the bandwidth of the filterbank as the frequency difference (in octaves) at which the bin has fallen to $\exp(-1/2)$ of its peak gain. $f_{\min}$ is the frequency of the lowest bin ($\ell = 0$) and N_O is the number of bins per octave in the log-frequency axis. The calculation is illustrated in Fig. 3, where the top-left image is the matrix $\mathbf{M}$, the top right is the conventional spectrogram $\mathbf{X}$, and the bottom right shows the resulting log-frequency spectrogram $\mathbf{Y}$.

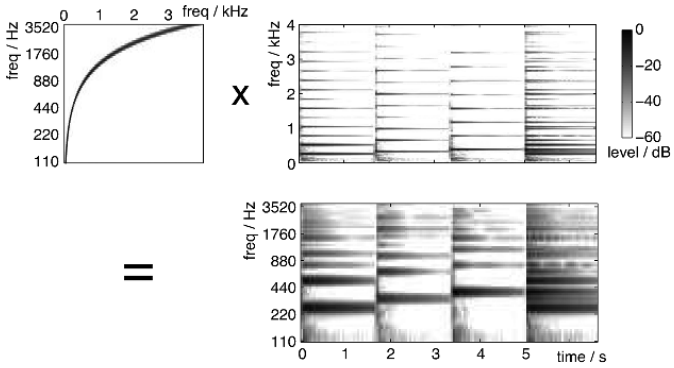

Fig. 3. Calculation of a log-frequency spectrogram as a columnwise linear mapping of bins from a conventional (linear-frequency) spectrogram. The images may be interpreted as matrices, but note the bottom-to-top sense of the frequency (row) axis.

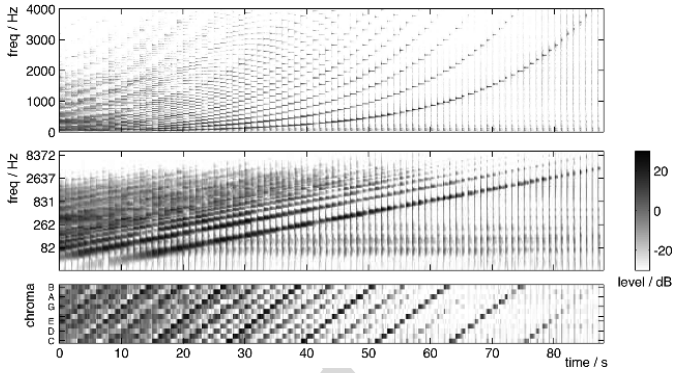

Fig. 4. Three representations of a chromatic scale comprising every note on the piano from lowest to highest. Top pane: conventional spectrogram (93-ms window). Middle pane: log-frequency spectrogram (186-ms window). Bottom pane: chromagram (based on 186-ms window).

Although conceptually simple, such a mapping often gives unsatisfactory results: in the figure, the logarithmic frequency axis uses $N_O = 12$ (one bin per semitone), starting at $f_{\min} = 110$ Hz (A2). At this point, the log-frequency bins have centers only 6.5 Hz apart; to have these centered on distinct STFT bins would require a window of 153 ms, or almost 7000 points at $F_s = 44\,100$ Hz. Using a 64-ms window, as in the figure, causes blurring of the low-frequency bins. Yet, by the same token, the highest bins shown—five octaves above the lowest—involve averaging together many STFT bins.

The long time window required to achieve semitone resolution at low frequencies has serious implications for the temporal resolution of any analysis. Since human perception of rhythm can often discriminate changes of 10 ms or less [4], an analysis window of 100 ms or more can lose important temporal structure. One popular alternative to a single STFT analysis is to construct a bank of individual bandpass filters, for instance one per semitone, each tuned the appropriate bandwidth and with minimal temporal support [6, Sec. 3.1]. Although this loses the famed computational efficiency of the fast Fourier transform, some of this may be regained by processing the highest octave with an STFT-based method, downsampling by a factor of 2, then repeating for as many octaves as are desired [7], [8]. This results in different sampling rates for each octave of the analysis, raising further computational issues. A toolkit for such analysis has been created by Schorkhuber and Klapuri [9].

E. Time-Chroma Representations

Some applications are primarily concerned with the chroma of the notes present, but less with the octave. Foremost among these is chord transcription—the annotation of the current chord as it changes through a song. Chords are a joint property of all the notes sounding at or near a particular point in time, for instance the C Major chord of Fig. 2, which is the unambiguous label of the three notes C, E, and G. Chords are generally defined by three or four notes, but the precise octave in which those notes occur is of secondary importance. Thus, for chord recognition, a representation that describes the chroma present but "folds" the octaves together seems ideal. This is the intention of chroma representations, first introduced as Pitch Class Profiles in [10]; the description as chroma was introduced in [11].

A typical chroma representation consists of a 12-bin vector for each time step, one for each chroma class from C to B. Given a log-frequency spectrogram representation with semitone resolution from the preceding section, one way to create chroma vectors is simply to add together all the bins corresponding to each distinct chroma [6, Ch. 3]. More involved approaches may include efforts to include energy only from strong sinusoidal components in the audio, and exclude non-tonal energy such as percussion and other noise. Estimating the precise frequency of tones in the lower frequency range may be important if the frequency binning of underlying transform is not precisely aligned to the musical scale [12].

Fig. 4 shows a chromatic scale, consisting of all 88 piano keys played one a second in an ascending sequence. The top pane shows the conventional, linear-frequency spectrogram, and the middle pane shows a log-frequency spectrogram calculated as in Fig. 3. Notice how the constant ratio between the fundamental frequencies of successive notes appears as an exponential growth on a linear axis, but becomes a straight line on a logarithmic axis. The bottom pane shows a 12-bin chroma representation (a "chromagram") of the same data. Even though there is only one note sounding at each time, notice that very few notes result in a chroma vector with energy in only a single bin. This is because although the fundamental may be mapped neatly into the appropriate chroma bin, as will the harmonics at $2f_0, 4f_0, 8f_0$, etc. (all related to the fundamental by octaves), the other harmonics will map onto other chroma bins. The harmonic at $3f_0$, for instance, corresponds to an octave plus 7 semitones ($2^{(12+7)/12} \approx 3$), thus for the C4 sounding at 40 s, we see the second most intense chroma bin after C is the G seven steps higher. Other harmonics fall in other bins, giving the more complex pattern. Many musical notes have the highest energy in the fundamental harmonic, and even with a weak fundamental, the root chroma is the bin into which the greatest number of low-order harmonics fall, but for a note with energy across a large number of harmonics—such as the lowest notes in the figure—the chroma vector can become quite cluttered.

One might think that attempting to attenuate higher harmonics would give better chroma representations by reducing these alias terms. In fact, many applications are improved by whitening the spectrum—i.e., boosting weaker bands to make the energy approximately constant across the spectrum. This

helps remove differences arising from the different spectral balance of different musical instruments, and hence better represents the tonal, and not the timbral or instrument-dependent, content of the audio. In [13], this is achieved by explicit normalization within a sliding local window, whereas [14] discards low-order cepstral coefficients as a form of "liftering."

Chroma representations may use more than 12 bins per octave to reflect finer pitch variations, but still retain the property of combining energy from frequencies separated by an octave [15], [16]. To obtain robustness against global mistunings (resulting from instruments tuned to a standard other than the 440 Hz A4, or distorted through equipment such as tape recorders running at the wrong speed), practical chroma analyses need to employ some kind of adaptive tuning, for instance by building a histogram of the differences between the frequencies of all strong harmonics and the nearest quantized semitone frequency, then shifting the semitone grid to match the peak of this histogram [12]. It is, however, useful to limit the range of frequencies over which chroma is calculated. Human pitch perception is most strongly influenced by harmonics that occur in a "dominance region" between about 400 and 2000 Hz [4]. Thus, after whitening, the harmonics can be shaped by a smooth, tapered frequency window to favor this range.

Notwithstanding the claims above that octave is less important than chroma, the lowest pitch in a collection of notes has a particularly important role in shaping the perception of simultaneous notes. This is why many musical ensembles feature a "bass" instrument—the double-bass in an orchestra, or the bass guitar in rock music—responsible for playing very low notes. As discussed in Section V, some applications explicitly track a bass line, but this hard decision can be avoided be avoided by calculating a second chroma vector over a lower frequency window, for instance covering 50 Hz to 400 Hz [17].

Code toolboxes to calculate chroma features are provided by Ellis[2] and Müller.[3]

F. Example Applications

Tonal representations—especially chroma features—have been used for a wide range of music analysis and retrieval tasks in which it is more important to capture polyphonic musical content without necessarily being concerned about the instrumentation. Such applications include chord recognition, alignment, "cover song" detection, and structure analysis.

1) Chord Recognition: As discussed above, chroma features were introduced as Pitch Class Profiles in [10], specifically for the task of recognizing chords in music audio. As a global property of the current notes and context, chords can be recognized based on a global representation of a short window. Moreover, shifting notes up or down by an octave rarely has much impact on the chord identity, so the octave-invariant properties of the chroma vector make it particularly appropriate. There has been a large amount of subsequent work on chord recognition, all based on some variant of chroma [15], [18]–[22]. Developments have mainly focused on the learning and classification aspects of the system: for instance, [15] noted the direct analogy between music audio with transcribed chord sequences (e.g., in "real books") that lack exact temporal alignments, and the speech recordings with unaligned word transcriptions used to train speech recognitions: they used the same Baum–Welch procedure to simultaneously estimate both models for the features of each chord, and the label alignment in the training data. Although later work has been able to take advantage of an increasing volume of manually labeled chord transcriptions [23], significant benefits have been attributed to refinements in feature extraction including separation of tonal (sinusoidal) and percussive (transient or noisy) energy [22], and local spectral normalization prior to chroma calculation [13].

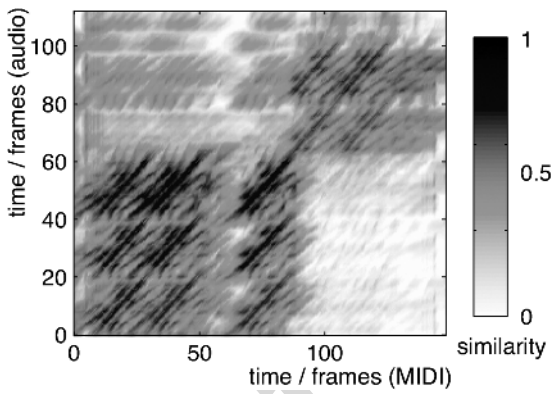

Fig. 5. Similarity matrix comparing a MIDI version of "And I Love Her" (horizontal axis) with the original Beatles recording (vertical axis). From [24].

2) Synchronization and Alignment: A difficult task such as chord recognition or polyphonic note transcription can be made substantially easier by employing an existing, symbolic description such as a known chord sequence, or the entire musical score. Then, the problem becomes that of *aligning* or *synchronizing* the symbolic description to the music audio [6, Ch. 5], making possible innovative applications such as an animated musical score display that synchronously highlights the appropriate bar in time with the music [25]. The core of such an application is to align compatible representations of each component with an efficient technique such as Dynamic Time Warping (DTW) [26], [27]. A time-chroma representation can be directly predicted from the symbolic description, or an electronic score such as a MIDI file can be synthesized into audio, then the synthesized form itself analyzed for alignment to the original recording [28], [29]. Fig. 5 shows a similarity matrix comparing a MIDI version of "And I Love Her" by the Beatles with the actual recording [24]. A similarity matrix $\mathbf{M_{sim}}(t_a, t_b)$ is populated by some measure of similarity $S(\mathbf{X_a}(t_a), \mathbf{X_b}(t_b))$ between representation $\mathbf{X_a}$ at time t_a and $\mathbf{X_b}$ at t_b. For instance, if both $\mathbf{X_a}$ and $\mathbf{X_b}$ are chroma representations, S could be the normalized correlation, $S(\mathbf{X}, \mathbf{Y}) = \mathbf{X} \cdot \mathbf{Y}/(|\mathbf{X}||\mathbf{Y}|)$. Differences in structure and timing between the two versions are revealed by wiggles in the dark ridge closest to the leading diagonal; flanking ridges relate to musical structure, discussed below.

A similar approach can synchronize different recordings of the same piece of music, for instance to allow switching, in real-time, between performances of a piano sonata by different

[2][Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/.

[3][Online]. Available: http://www.mpi-inf.mpg.de/~mmueller/chromatoolbox/.

pianists [25], [30], [31]. In this case, the relatively poor temporal accuracy of tonal representations may be enhanced by the addition of features better able to achieve precise synchronization of onset events [32].

3) "Cover Song" Detection: In popular music, an artist may record his or her own version of another artist's composition, often incorporating substantial changes to instrumentation, tempo, structure, and other stylistic aspects. These alternate interpretations are sometimes known as "cover" versions, and present a greater challenge to alignment, due to the substantial changes. The techniques of chroma feature representation and DTW are, however, still the dominant approaches over several years of development and formal evaluation of this task within the MIREX campaign [33]. Gross structural changes will interfere with conventional DTW, so it must either be modified to report "local matches" [34], or replaced by a different technique such as cross-correlation of beat-synchronous chroma representations [12]. The need for efficient search within very large music collections can be satisfied with efficient hash-table representation of the music broken up into smaller fragments [35]–[37].

4) Structure Recovery: A similarity matrix that compares a piece of music to itself will have a perfectly straight leading diagonal ridge, but will likely have flanking ridges similar to those visible in Fig. 5. These ridges indicate that a certain portion of the signal resembles an earlier (or later) part—i.e., the signal exhibits some repetitive structure. Recurring melodies and chord sequences are ubiquitous in music, which frequently exhibits a hierarchical structure. In popular music, for instance, the song may consist of an introduction, a sequence of alternating verse and chorus, a solo or bridge section, etc. Each of these segments may in turn consist of multiple phrases or lines with related or repeating structure, and the individual phrases may themselves consist of repeating or nearly repeating patterns of notes. [38], for instance, argues that the observation and acquisition of this kind of structure is an important part of the enjoyment of music listening.

Automatic segmentation and decomposition according to this structure is receiving an increasing level of attention; see [39] for a recent review. Typically, systems operate by 1) finding off-diagonal ridges in a similarity matrix to identify and segment into repeating phrases [40], [41], and/or 2) finding segmentation points such that some measure of statistical similarity is maximized within segments, but minimized between adjacent segments [42], [43]. Since human labelers exhibit less consistency on this annotation tasks than for, say, beat or chord labeling, structure recovery is sometimes simplified into problems such as identifying the "chorus," a frequently repeated and usually obvious part of popular songs [11]. Other related problems include searching for structures and motifs that recur within and across different songs within a given body of music [44], [45].

III. Tempo, Beat, and Rhythm

The musical aspects of tempo, beat, and rhythm play a fundamental role for the understanding of and the interaction with music [46]. It is the *beat*, the steady pulse that drives music forward and provides the temporal framework of a piece of music

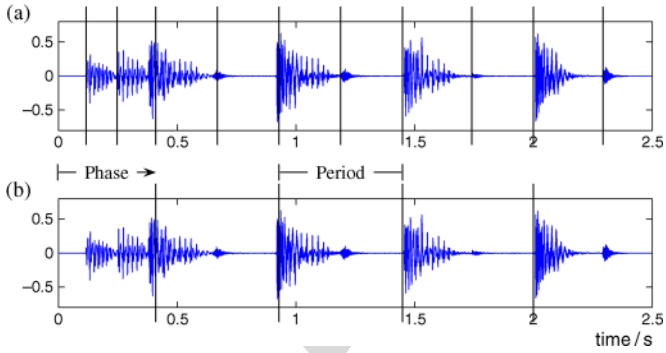

Fig. 6. Waveform representation of the beginning of *Another one bites the dust* by Queen. (a) Note onsets. (b) Beat positions.

[47]. Intuitively, the beat can be described as a sequence of perceived pulses that are regularly spaced in time and correspond to the pulse a human taps along when listening to the music [48]. The term *tempo* then refers to the rate of the pulse. Musical pulses typically go along with note onsets or percussive events. Locating such events within a given signal constitutes a fundamental task, which is often referred to as *onset detection*. In this section, we give an overview of recent approaches for extracting onset, tempo, and beat information from music signals, and then indicate how this information can be applied to derive higher-level rhythmic patterns.

A. Onset Detection and Novelty Curve

The objective of *onset detection* is to determine the physical starting times of notes or other musical events as they occur in a music recording. The general idea is to capture sudden changes in the music signal, which are typically caused by the onset of novel events. As a result, one obtains a so-called *novelty curve*, the peaks of which indicate onset candidates. Many different methods for computing novelty curves have been proposed; see [49] and [50] for an overview. For example, playing a note on a percussive instrument typically results in a sudden increase of the signal's energy, see Fig. 6(a). Having such a pronounced attack phase, note onset candidates may be determined by locating time positions, where the signal's amplitude envelope starts to increase [49]. Much more challenging, however, is the detection of onsets in the case of non-percussive music, where one often has to deal with soft onsets or blurred note transitions. This is often the case for vocal music or classical music dominated by string instruments. Furthermore, in complex polyphonic mixtures, simultaneously occurring events may result in masking effects, which makes it hard to detect individual onsets. As a consequence, more refined methods have to be used for computing the novelty curves, e.g., by analyzing the signal's spectral content [49], [51], pitch [51], [52], harmony [53], [54], or phase [49], [55]. To handle the variety of different signal types, a combination of novelty curves particularly designed for certain classes of instruments can improve the detection accuracy [51], [56]. Furthermore, to resolve masking effects, detection functions were proposed that analyze the signal in a bandwise fashion to extract transients occurring in certain frequency regions of the signal [57], [58]. For example, as a side-effect of

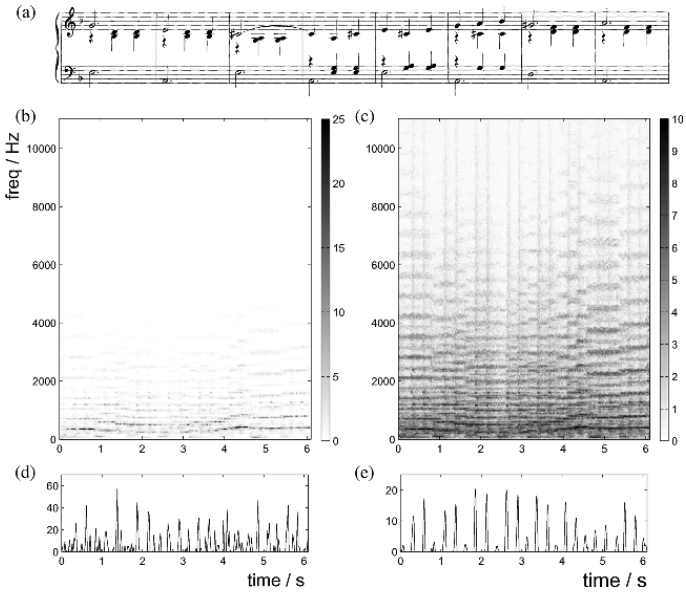

Fig. 7. Excerpt of Shostakovich's Waltz No. 2 from the *Suite for Variety Orchestra No.* 1. (a) Score representation (in a piano reduced version). (b) Magnitude spectrogram. (c) Compressed spectrogram using $C = 1000$. (d) Novelty curve derived from (b). (e) Novelty curve derived from (c).

a sudden energy increase, one can often observe an accompanying broadband noise burst in the signal's spectrum. This effect is mostly masked by the signal's energy in lower frequency regions, but it is well detectable in the higher frequency regions of the spectrum [59]. A widely used approach to onset detection in the frequency domain is the *spectral flux* [49], [60], where changes of pitch and timbre are detected by analyzing the signal's short-time spectrum.

To illustrate some of these ideas, we now describe a typical spectral-based approach for computing novelty curves. Given a music recording, a short-time Fourier transform is used to obtain a spectrogram $X = (X(t,k))_{t,k}$ with $k \in [0 : K]$ and $t \in [0 : T-1]$ as in (1). Note that the Fourier coefficients of X are linearly spaced on the frequency axis. Using suitable binning strategies, various approaches switch over to a logarithmically spaced frequency axis, e.g., by using mel-frequency bands or pitch bands; see [57], [58], and Section II-D. Keeping the linear frequency axis puts greater emphasis on the high-frequency regions of the signal, thus accentuating the aforementioned noise bursts visible as high-frequency content. One simple, yet important step, which is often applied in the processing of music signals, is referred to as *logarithmic compression*; see [57]. In our context, this step consists in applying a logarithm to the magnitude spectrogram $|X|$ of the signal yielding $Y = \log(1+C\cdot|X|)$ for a suitable constant $C > 1$. Such a compression step not only accounts for the logarithmic sensation of human sound intensity, but also balances out the dynamic range of the signal. In particular, by increasing C, low-intensity values in the high-frequency spectrum become more prominent. This effect is clearly visible in Fig. 7, which shows the magnitude spectrogram $|X|$ and the compressed spectrogram Y for a recording of a Waltz by Shostakovich. On the downside, a large compression factor C may also amplify non-relevant low-energy noise components.

To obtain a novelty curve, one basically computes the discrete derivative of the compressed spectrum Y. More precisely, one sums up only positive intensity changes to emphasize onsets while discarding offsets to obtain the novelty function $\Delta : [0 : T-2] \to \mathbb{R}$:

$$\Delta(t) = \sum_{k=0}^{K} |Y(t+1,k) - Y(t,k)|_{\geq 0} \tag{4}$$

for $t \in [0 : T-2]$, where $|x|_{\geq 0} = x$ for a non-negative real number x and $|x|_{\geq 0} = 0$ for a negative real number x. In many implementations, higher order smoothed differentiators are used [61] and the resulting curve is further normalized [62], [63]. Fig. 7(e) shows a typical novelty curve for our Shostakovich example. As mentioned above, one often process the spectrum in a bandwise fashion obtaining a novelty curve for each band separately [57], [58]. These novelty curves are then weighted and summed up to yield a final novelty function.

The peaks of the novelty curve typically indicate the positions of note onsets. Therefore, to explicitly determine the positions of note onsets, one employs peak picking strategies based on fixed or adaptive thresholding [49], [51]. In the case of noisy novelty curves with many spurious peaks, however, this is a fragile and error-prone step. Here, the selection of the relevant peaks that correspond to true note onsets becomes a difficult or even infeasible problem. For example, in the Shostakovich Waltz, the first beats (downbeats) of the 3/4 meter are played softly by non-percussive instruments leading to relatively weak and blurred onsets, whereas the second and third beats are played staccato supported by percussive instruments. As a result, the peaks of the novelty curve corresponding to downbeats are hardly visible or even missing, whereas peaks corresponding to the percussive beats are much more pronounced, see Fig. 7(e).

B. Periodicity Analysis and Tempo Estimation

Avoiding the explicit determination of note onset, novelty curves are often directly analyzed in order to detect reoccurring or quasi-periodic patterns, see [64] for an overview of various approaches. Here, generally speaking, one can distinguish between three different methods. The autocorrelation method allows for detecting periodic self-similarities by comparing a novelty curve with time-shifted (localized) copies [65]–[68]. Another widely used method is based on a bank of comb filter resonators, where a novelty curve is compared with templates that consists of equally spaced spikes covering a range of periods and phases [57], [58]. Third, the short-time Fourier transform can be used to derive a time–frequency representation of the novelty curve [62], [63], [67]. Here, the novelty curve is compared with templates consisting of sinusoidal kernels each representing a specific frequency. Each of the methods reveals periodicity properties of the underlying novelty curve from which one can estimate the tempo or beat structure. The intensities of the estimated periodicity, tempo, or beat properties typically change over time and are often visualized by means of spectrogram-like representations referred to as *tempogram* [69], *rhythmogram* [70], or *beat spectrogram* [71].

Exemplarily, we introduce the concept of a tempogram while discussing two different periodicity estimation methods. Let $[0 : T-1]$ (as for the novelty curve) denote the sampled time axis,

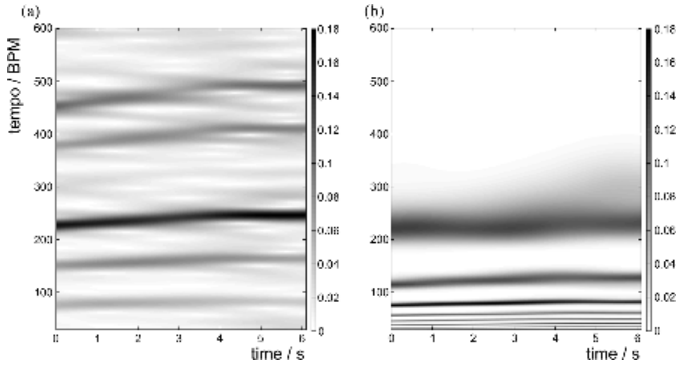

Fig. 8. Excerpt of Shostakovich's Waltz No. 2 from the *Suite for Variety Orchestra No.* 1. (a) Fourier tempogram. (b) Autocorrelation tempogram.

which we extend to $\mathbb{Z}$ to avoid boundary problems. Furthermore, let $\Theta \subset \mathbb{R}_{>0}$ be a set of tempi specified in beats per minute (BPM). Then, a tempogram is mapping $\mathcal{T} : \mathbb{Z} \times \Theta \to \mathbb{R}_{\geq 0}$ yielding a time-tempo representation for a given time-dependent signal. For example, suppose that a music signal has a dominant tempo of $\tau = 220$ BPM around position t, then the corresponding value $\mathcal{T}(t,\tau)$ is large, see Fig. 8. In practice, one often has to deal with tempo ambiguities, where a tempo τ is confused with integer multiples $\tau, 2\tau, 3\tau, \ldots$ (referred to as *harmonics* of τ) and integer fractions $\tau, \tau/2, \tau/3, \ldots$ (referred to as *subharmonics* of τ). To avoid such ambiguities, a mid-level tempo representation referred to as *cyclic tempograms* can be constructed, where tempi differing by a power of two are identified [72], [73]. This concept is similar to the cyclic chroma features, where pitches differing by octaves are identified, cf. Section II-E. We discuss the problem of tempo ambiguity and pulse level confusion in more detail in Section III-C.

A tempogram can be obtained by analyzing a novelty curve Δ with respect to local periodic patterns using a short-time Fourier transform [62], [63], [67]. To this end, one fixes a window function $W : \mathbb{Z} \to \mathbb{R}$ of finite length centered at $t = 0$ (e.g., a centered Hann window of size $2N+1$ for some $N \in \mathbb{N}$). Then, for a frequency parameter $\omega \in \mathbb{R}_{\geq 0}$, the complex Fourier coefficient $\mathcal{F}(t,\omega)$ is defined by

$$\mathcal{F}(t,\omega) = \sum_{n \in \mathbb{Z}} \Delta(n) \cdot W(n-t) \cdot \exp\{-2\pi j \omega n\}. \quad (5)$$

Note that the frequency parameter ω (measured in Hertz) corresponds to the tempo parameter $\tau = 60 \cdot \omega$ (measured in BPM). Therefore, one obtains a discrete *Fourier tempogram* $\mathcal{T}^{\mathrm{F}} : \mathbb{Z} \times \Theta \to \mathbb{R}_{\geq 0}$ by

$$\mathcal{T}^{\mathrm{F}}(t,\tau) = |\mathcal{F}(t,\tau/60)|. \quad (6)$$

As an example, Fig. 8(a) shows the tempogram $\mathcal{T}^{\mathrm{F}}$ of our Shostakovich example from Fig. 7. Note that $\mathcal{T}^{\mathrm{F}}$ reveals a slightly increasing tempo over time starting with roughly $\tau = 225$ BPM. Also, $\mathcal{T}^{\mathrm{F}}$ reveals the second tempo harmonics starting with $\tau = 450$ BPM. Actually, since the novelty curve Δ locally behaves like a track of positive clicks, it is not hard to see that Fourier analysis responds to harmonics but tends to suppress subharmonics, see also [73], [74].

Also autocorrelation-based methods are widely used to estimate local periodicities [66]. Since these methods, as it turns out, respond to subharmonics while suppressing harmonics, they ideally complement Fourier-based methods, see [73], [74]. To obtain a discrete *autocorrelation tempogram*, one again fixes a window function $W : \mathbb{Z} \to \mathbb{R}$ centered at $t = 0$ with support $[-N : N]$, $N \in \mathbb{N}$. The local autocorrelation is then computed by comparing the windowed novelty curve with time shifted copies of itself. Here, we use the unbiased local autocorrelation

$$\mathcal{A}(t,\ell) = \frac{\sum_{n \in \mathbb{Z}} \Delta(n) W(n-t) \Delta(n+\ell) \cdot W(n-t+\ell)}{2N+1-\ell} \quad (7)$$

for time $t \in \mathbb{Z}$ and time lag $\ell \in [0 : N]$. Now, to convert the lag parameter into a tempo parameter, one needs to know the sampling rate. Supposing that each time parameter $t \in \mathbb{Z}$ corresponds to r seconds, then the lag ℓ corresponds to the tempo $\tau = 60/(r \cdot \ell)$ BPM. From this, one obtains the *autocorrelation tempogram* $\mathcal{T}^{\mathrm{A}}$ by

$$\mathcal{T}^{\mathrm{A}}(t,\tau) = \mathcal{A}(t,\ell). \quad (8)$$

for each tempo $\tau = 60/(r \cdot \ell)$, $\ell \in [1 : N]$. Finally, using standard resampling and interpolation techniques applied to the tempo domain, one can derive an autocorrelation tempogram $\mathcal{T}^{\mathrm{A}} : \mathbb{Z} \times \Theta \to \mathbb{R}_{\geq 0}$ that is defined on the same tempo set $Theta$ as the Fourier tempogram $\mathcal{T}^{\mathrm{F}}$. The tempogram $\mathcal{T}^{\mathrm{A}}$ for our Shostakovich example is shown in Fig. 8(b), which clearly indicates the subharmonics. Actually, the parameter $\tau = 75$ is the third subharmonics of $\tau = 225$ and corresponds to the tempo on the measure level.

Assuming a more or less steady tempo, most tempo estimation approaches determine only one *global* tempo value for the entire recording. For example, such a value may be obtained by averaging the tempo values (e.g., using a median filter [53]) obtained from a framewise periodicity analysis. Dealing with music with significant tempo changes, the task of *local* tempo estimation (for each point in time) becomes a much more difficult or even ill-posed problem; see also Fig. 9 for a complex example. Having computed a tempogram, the framewise maximum yields a good indicator of the locally dominating tempo—however, one often has to struggle with confusions of tempo harmonics and subharmonics. Here, tempo estimation can be improved by a combined usage of Fourier and autocorrelation tempograms. Furthermore, instead of simply taking the framewise maximum, global optimization techniques based on dynamic programming have been suggested to obtain smooth tempo trajectories [61], [67].

C. Beat Tracking

When listening to a piece of music, most humans are able to tap to the musical beat without difficulty. However, transferring this cognitive process into an automated system that reliably works for the large variety of musical styles is a challenging task. In particular, the tracking of beat positions becomes hard in the case that a music recording reveals significant tempo changes. This typically occurs in expressive performances of classical music as a result of ritardandi, accelerandi, fermatas, and artistic shaping [75]. Furthermore, the extraction problem is complicated by the fact that there are various levels that are presumed to contribute to the human perception of beat.

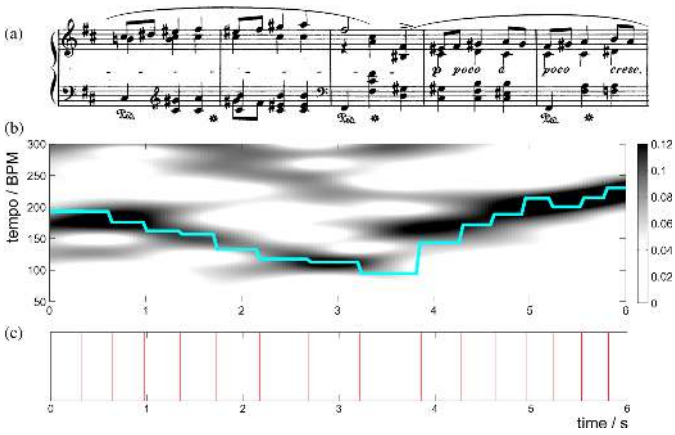

Fig. 9. Excerpt of the Mazurka Op. 30 No. 2 (played by Rubinstein, 1966). (a) Score. (b) Fourier tempogram with reference tempo (cyan). (c) Beat positions (quarter note level).

Most approaches focus on determining musical pulses on the *tactus* level (the foot tapping rate) [65]–[67], but only few approaches exist for analyzing the signal on the measure level [54], [57] or finer tatum level [76]–[78]. Here, a *tatum* or *temporal atom* refers to the fastest repetition rate of musically meaningful accents occurring in the signal [79]. Various approaches have been suggested that simultaneously analyze different pulse levels [57], [68], [80]. In [62] and [63], instead of looking at a specific pulse level, a robust mid-level representation has been introduced which captures the predominant local pulse even in the presence of significant tempo fluctuations.

Exemplarily, we describe a robust beat tracking procedure [66], which assumes a roughly constant tempo throughout the music recording. The input of the algorithm consists of a novelty curve $\Delta : [0 : T-1] \to \mathbb{R}$ as well as an estimate $\hat{\tau}$ of the global (average) tempo, which also determines the pulse level to be considered. From $\hat{\tau}$ and the sampling rate used for the novelty curve, one can derive an estimate $\hat{\rho} \in \mathbb{Z}$ for the average beat period (given in samples). Assuming a roughly constant tempo, the difference δ of two neighboring beats should be close to $\hat{\rho}$. To measure the distance between δ and $\hat{\rho}$, a neighborhood function $N_{\hat{\rho}} : \mathbb{N} \to \mathbb{R}$, $N_{\hat{\rho}}(\delta) = -(\log_2(\delta/\hat{\rho}))^2$, is introduced. This function takes the maximum value of 0 for $\delta = \hat{\rho}$ and is symmetric on a log-time axis. Now, the task is to estimate a sequence $B = (b_1, b_2, \ldots, b_L)$, for some suitable $L \in \mathbb{N}$, of monotonously increasing beat positions $b_\ell \in [0 : T-1]$ satisfying two conditions. On the one hand, the value $\Delta(b_\ell)$ should be large for all $\ell \in [1 : L]$, and, on the other hand, the beat intervals $\delta = b_\ell - b_{\ell-1}$ should be close to $\hat{\rho}$. To this end, one defines the score $S(B)$ of a beat sequence $B = (b_1, b_2, \ldots, b_L)$ by

$$S(B) = \sum_{\ell=1}^{L} \Delta(b_\ell) + \alpha \sum_{\ell=2}^{L} N_{\hat{\rho}}(b_\ell - b_{\ell-1}) \qquad (9)$$

where the weight $\alpha \in \mathbb{R}$ balances out the two conditions. Finally, the beat sequence maximizing S yields the solution of the beat tracking problem. The score-maximizing beat sequence can be obtained by a straightforward dynamic programming (DP) approach; see [66] for details.

As mentioned above, recent beat tracking procedures work well for modern pop and rock music with a strong and steady beat, but the extraction of beat locations from highly expressive performances still constitutes a challenging task with many open problems. For such music, one often has significant local tempo fluctuation caused by the artistic freedom a musician takes, so that the model assumption of local periodicity is strongly violated. This is illustrated by Fig. 9, which shows a tempo curve and the beat positions for a romantic piano music recording (Mazurka by Chopin). In practice beat tracking is further complicated by the fact that there may be beats with no explicit note events going along with them [81]. Here, a human may still perceive a steady beat by subconsciously interpolating the missing onsets. This is a hard task for a machine, in particular in passages of varying tempo where interpolation is not straightforward. Furthermore, auxiliary note onsets can cause difficulty or ambiguity in defining a specific physical beat time. In music such as the Chopin Mazurkas, the main melody is often embellished by ornamented notes such as trills, grace notes, or arpeggios. Also, for the sake of expressiveness, the notes of a chord need not be played at the same time, but slightly displaced in time. This renders a precise definition of a physical beat position impossible [82]. Such highly expressive music also reveals the limits of purely onset-oriented tempo and beat tracking procedures, see also [63].

D. Higher-Level Rhythmic Structures

The extraction of onset, beat, and tempo information is of fundamental importance for the determination of higher-level musical structures such as rhythm and meter [46], [48]. Generally, the term *rhythm* is used to refer to a temporal patterning of event durations, which are determined by a regular succession of strong and weak stimuli [83]. Furthermore, the perception of rhythmic patterns also depends on other cues such as the dynamics and timbre of the involved sound events. Such repeating patterns of accents form characteristic *pulse groups*, which determine the *meter* of a piece of music. Here, each group typically starts with an accented beat and consists of all pulses until the next accent. In this sense, the term *meter* is often used synonymously with the term *time signature*, which specifies the beat structure of a musical measure or bar. It expresses a regular pattern of beat stresses continuing through a piece thus defining a hierarchical grid of beats at various time scales.

Rhythm and tempo are often sufficient for characterizing the style of a piece of music. This particularly holds for dance music, where, e.g., a waltz or tango can be instantly recognized from the underlying rhythmic pattern. Various approaches have been described for determining some kind of rhythm template, which have mainly been applied for music classification tasks [77], [84], [85]. Typically, the first step consists in performing some beat tracking. In the next step, assuming additional knowledge such as the time signature and the starting position of the first bar, patterns of alternating strong and weak pulses are determined for each bar, which are then averaged over all bars to yield an average rhythmic pattern for the entire piece [84], [85]. Even though such patterns may still be abstract, they have been successfully applied for tasks such as dance style classification. The automatic extraction of explicit rhythmic

parameters such as the time signature constitutes a difficult problem. A first step towards time signature estimation has been described in [86], where the number of beats between regularly recurring accents (or downbeats) are estimated to distinguish between music having a duple or triple meter.

Another way for deriving rhythm-related features is to consider intervals defined by successive onset or beat positions, often referred as *inter-onset-intervals* (IOIs). Considering histograms over the durations of occurring IOIs, one may then derive hypotheses on the beat period, tempo, and meter [75], [76], [87]. The drawback of these approaches is that they rely on an explicit localization of a discrete set of onset and beat positions—a fragile and error-prone step. To compensate for such errors, various approaches have been proposed to jointly or iteratively estimate onset, pulse, and meter parameters [54], [78].

IV. Timbre and Instrumentation

Timbre is defined as the "attribute of auditory sensation in terms of which a listener can judge two sounds similarly presented and having the same loudness and pitch as dissimilar" [88]. The concept is closely related to sound source recognition: for example, the sounds of the violin and the flute may be identical in their pitch and loudness, but are still easily distinguished. Furthermore, when listening to polyphonic music, we are usually able to perceptually organize the component sounds to their sources based on timbre information.

The term *polyphonic timbre* refers to the overall timbral mixture of a music signal, the "global sound" of a piece of music [89], [90]. Human listeners, especially trained musicians, can switch between a "holistic" listening mode where they consider a music signal as a coherent whole, and a more analytic mode where they focus on the part played by a particular instrument [91], [92]. In computational systems, acoustic features describing the polyphonic timbre have been found effective for tasks such as automatic genre identification [93], music emotion recognition [94], and automatic tagging of audio with semantic descriptors [95]. A computational analogy for the analytical listening mode, in turn, includes recognizing musical instruments on polyphonic recordings.

This section will first discuss feature representations for timbre and then review methods for musical instrument recognition in isolation and in polyphonic music signals.

A. Perceptual Dimensions of Timbre

Timbre is a multidimensional concept, having several underlying acoustic factors. Schouten [96] describes timbre as being determined by five major acoustic parameters: 1) the range between tonal and noise-like character; 2) the spectral envelope; 3) the time envelope; 4) the changes of spectral envelope and fundamental frequency; and 5) the onset of the sound differing notably from the sustained vibration.

The perceptual dimensions of timbre have been studied based on dissimilarity ratings of human listeners for sound pairs; see [97] and [98]. In these studies, multidimensional scaling (MDS) was used to project the dissimilarity ratings into a lower-dimensional space where the distances between the sounds match as closely as possible the dissimilarity ratings. Acoustic correlates can then be proposed for each dimension of this timbre space. Several studies report spectral centroid, $c_t = \sum_k |kX(t,k)| / \sum_k |X(t,k)|$, and attack time as major determinants of timbre. Also often reported are spectral irregularity (defined as the average level difference of neighboring harmonics) and spectral flux, $d = (1/T)\sum_t ||\mathbf{d}_t||$, where $\mathbf{d}_t = |X(t,k)| - |X(t-1,k)|$.

Very few studies have attempted to uncover the perceptual dimensions of polyphonic timbre. Cogan [99] carried out informal musicological case studies using the spectrograms of diverse music signals and proposed 13 dimensions to describe the quality of musical sounds. Furthermore, Kendall and Carterette [100] studied the perceptual dimensions of simultaneous wind instrument timbres using MDS, whereas Alluri and Toiviainen [90] explored the polyphonic timbre of Indian popular music. The latter observed relatively high correlations between certain perceptual dimensions and acoustic features describing spectro-temporal modulations.

B. Time-Varying Spectral Envelope

The acoustic features found in the MDS experiments bring insight into timbre perception, but they are generally too low-dimensional to lead to robust musical instrument identification [101]. In signal processing applications, timbre is typically described using a parametric model of the time-varying spectral envelope of sounds. This stems from speech recognition [102] and is not completely satisfactory in music processing as will be seen in Section IV-C, but works well as a first approximation of timbre. Fig. 10 illustrates the time-varying spectral envelopes of two example musical tones. Indeed, all the acoustic features found in the MDS experiments are implicitly represented by the spectral envelope, and among the five points on Schouten's list, 2)–5) are reasonably well covered. The first point, tonal versus noiselike character, can be addressed by decomposing a music signal into its sinusoidal and stochastic components [103], [104], and then estimating the spectral envelope of each part separately. This, for example, has been found to significantly improve the accuracy of genre classification [105].

Mel-frequency cepstral coefficients (MFCCs), originally used for speech recognition [102], are by far the most popular way of describing the spectral envelope within an individual analysis frame. MFCCs encode the coarse shape of the log-power spectrum on the mel-frequency scale.[4] They have the desirable property that a small (resp. large) numerical change in the MFCC coefficients corresponds to a small (resp. large) perceptual change. MFCCs are calculated by simulating a bank of about 40 bandpass filters in the frequency domain (the filters being uniformly spaced on the Mel-frequency scale), calculating the log-power of the signal within each band, and finally applying a discrete cosine transform to the vector of log-powers to obtain the MFCC coefficients. Typically only the 10–15 lowest coefficients are retained and the rest are discarded in order to make the timbre features invariant to pitch information that is present in the higher coefficients. Time-varying aspects are usually accounted for by appending temporal derivatives of the MFCCs to the feature vector.

[4]The mel-frequency scale is one among several scales that model the frequency resolution of the human auditory system. See [106] for a comparison of different perceptual frequency scales.

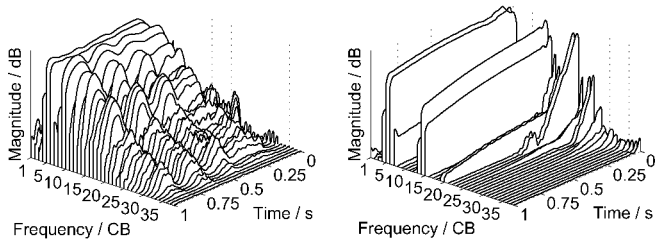

Fig. 10. Time-varying spectral envelopes of 260-Hz tones of the flute (left) and the vibraphone (right). Here sound pressure level within auditory critical bands (CB) is shown as a function of time.

Modulation spectrum encodes the temporal variation of spectral energy explicitly [107]. This representation is obtained by first using a filterbank to decompose an audio signal into subbands, extracting the energy envelope within each band, and finally analyzing amplitude modulations (AM) within each band by computing discrete Fourier transforms of the energy envelope within longer "texture" windows (phases are discarded to achieve shift-invariance). This results in a three-dimensional representation where the dimensions correspond to time, frequency, and AM frequency (typically in the range 0–200 Hz). Sometimes the time dimension can be collapsed by analyzing AM modulation in a single texture window covering the entire signal. Spectro-temporal modulations play an important role in the perception of polyphonic timbre [90]; therefore, representations based on modulation spectra are particularly suitable for describing the instrumentation aspects of complex music signals. Indeed, state-of-the-art genre classification is based on the modulation spectrum [108]. Other applications of the modulation spectrum include speech recognition [109], audio coding [107], and musical instrument recognition [110].

C. Source-Filter Model of Sound Production

Let us now consider more structured models of musical timbre. Instrument acoustics provides a rich source of information for constructing models for the purpose of instrument recognition. The source-filter model of sound production is particularly relevant here [111]. Many musical instruments can be viewed as a coupling of a vibrating object, such as a guitar string ("source"), with the resonance structure of the rest of the instrument ("filter") that colors the produced sound. The source part usually determines pitch, but often contains also timbral information.

The source-filter model has been successfully used in speech processing for decades [112]. However, an important difference between speech and music is that there is only one sound production mechanism in speech, whereas in music a wide variety of sound production mechanisms are employed. Depending on the instrument, the sound can be produced for example by vibrating strings, air columns, or vibrating bars, and therefore the source excitation provides valuable information about the instrument identity.

It is interesting to note that the regularities in the source excitation are not best described in terms of frequency, but in terms of harmonic index. For example the sound of the clarinet is characterized by the odd harmonics being stronger than the even harmonics. For the piano, every Mth partial is weaker because the string is excited at a point $1/M$ along its length. The sound of the vibraphone, in turn, exhibits mainly the first and the fourth harmonic and some energy around the tenth partial. MFCCs and other models that describe the properties of an instrument as a function of frequency smear out this information. Instead, a structured model is needed where the spectral information is described both as a function of frequency and as a function of harmonic index.

The source-filter model for the magnitude spectrum $|X(f)|$ of a harmonic sound can be written as

$$|X(f_h)| = \gamma S(h) B(f_h) \tag{10}$$

where $f_h \approx hv$, $h = 1, 2, \ldots$ is the frequency of the hth harmonic of a sound with fundamental frequency v. Note that $|X(f)|$ is modeled only at the positions of the harmonics and is assumed zero elsewhere. The scalar γ denotes the overall gain of the sound, $S(h)$ is the amplitude of harmonic h in the spectrum of the vibrating source, and $B(f)$ represents the frequency response of the instrument body. Perceptually, it makes sense to minimize the modeling error on the log-magnitude scale and therefore to take the logarithm of both sides of (10). This renders the model linear and allows the two parts, $\log S(h)$ and $\log B(f_h)$, to be further represented using a suitable linear basis [113]. In addition to speech coding and music synthesis [111], [112], the source-filter model has been used to separate the main melody from polyphonic music [114] and to recognize instruments in polyphonic music [115].

Above we assumed that the source excitation produces a spectrum where partial frequencies obey $f_h \approx hv$. Although such sounds are the commonplace in Western music (mallet percussion instruments being the exception), this is not the case in all music cultures and the effect of partial frequencies on timbre has been very little studied. Sethares [116] has investigated the relationship between the spectral structure of musical sounds and the structure of musical scales used in different cultures.

D. Recognition of Musical Instruments in Isolation

This section reviews techniques for musical instrument recognition in signals where only one instrument is playing at a time. Systems developed for this purpose typically employ the supervised classification paradigm (see Fig. 11), where 1) acoustic features are extracted in successive time frames in order to describe the relevant aspects of the signal; 2) training data representing each instrument class is used to learn a model for within-class feature distributions; and 3) the models are then used to classify previously unseen samples.

A number of different supervised classification methods have been used for instrument recognition, including k-nearest neighbors, Gaussian mixture models, hidden Markov models, linear discriminant analysis, artificial neural networks, support vector machines, and decision trees [117], [118]. For a comprehensive review of the different recognition systems for isolated notes and solo phrases, see [119]–[122].

A variety of acoustic features have been used for instrument recognition. Spectral features include the first few moments of the magnitude spectrum (spectral centroid, spread, skewness,

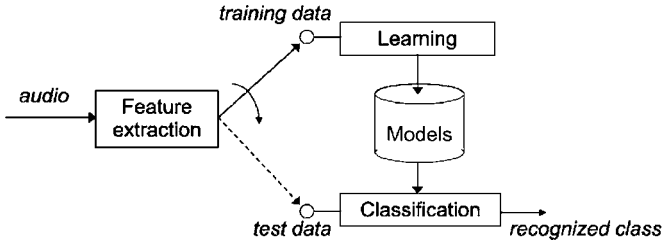

Fig. 11. General overview of supervised classification. See text for details.

and kurtosis), sub-band energies, spectral flux, spectral irregularity, and harmonic versus noise part energy [123]. Cepstral features include MFCCs and warped linear prediction-based cepstral coefficients, see [101] for a comparison. Modulation spectra have been used in [110]. Temporal features include the first few moments of the energy envelope within frames of about one second in length, and the frequency and strength of amplitude modulation in the range 4–8 Hz ("tremolo") and 10–40 Hz ("roughness") [122], [124]. The first and second temporal derivatives of the features are often appended to the features vector. For a more comprehensive list of acoustic features and comparative evaluations, see [101], [122], [125], [126].

Obviously, the above list of acoustic features is highly redundant. Development of a classification system typically involves a feature selection stage, where training data is used to identify and discard unnecessary features and thereby reduce the computational load of the feature extraction, see Herrera *et al.* [119] for a discussion on feature selection methods. In order to facilitate the subsequent statistical modeling of the feature distributions, the retained features are often decorrelated and the dimensionality of the feature vector is reduced using principal component analysis, linear discriminant analysis, or independent component analysis [117].

Most instrument classification systems resort to the so-called bag-of-features approach where an audio signal is modeled by the statistical distribution of its short-term acoustic features, and the temporal order of the features is ignored. An exception here are the instrument recognizers employing hidden Markov models where temporal dependencies are taken into account explicitly [127], [128]. Joder *et al.* [122] carried out an extensive evaluation of different temporal integration mechanisms to see if they improve over the bag-of-features approach. They found that a combination of feature-level and classifier-level temporal integration improved over a baseline system, although neither of them alone brought a significant advantage. Furthermore, HMMs performed better than GMMs, which suggests that taking into account the temporal dependencies of the feature vectors improves classification.

The techniques discussed above are directly applicable to other audio signal classification tasks too, including genre classification [93], automatic tagging of audio [95], and music emotion recognition [94], for example. However, the optimal acoustic features and models are usually specific to each task.

E. Instrument Recognition in Polyphonic Mixtures

Instrument recognition in polyphonic music is closely related to sound source separation: recognizing instruments in a mixture allows one to generate time–frequency masks that indicate which spectral components belong to which instrument. Vice versa, if individual instruments can be reliably separated from the mixture, the problem reduces to that of single-instrument recognition. The problem of source separation will be discussed in Section V.

A number of different approaches have been proposed for recognizing instruments in polyphonic music. These include extracting acoustic features directly from the mixture signal, sound source separation followed by the classification of each separated signal, signal model-based probabilistic inference, and dictionary-based methods. Each of these will be discussed in the following.

The most straightforward approach to polyphonic instrument recognition is to extract features directly from the mixture signal. Little and Pardo [129] used binary classifiers to detect the presence of individual instruments in polyphonic audio. They trained classifiers using weakly labeled mixture signals, meaning that only the presence or absence of the target sound object was indicated but not the exact times when it was active. They found that learning from weakly labeled mixtures led to better results than training with isolated examples of the target instrument. This was interpreted to be due to the fact that the training data, in the mixed case, was more representative of the polyphonic data on which the system was tested. Essid *et al.* [124] developed a system for recognizing *combinations* of instruments directly. Their method exploits hierarchical classification and an automatically built taxonomy of musical ensembles in order to represent every possible combination of instruments that is likely to be played simultaneously in a given genre.

Eggink and Brown [130] introduced missing feature theory to instrument recognition. Here the idea is to estimate a binary mask that indicates time–frequency regions that are dominated by energy from interfering sounds and are therefore to be excluded from the classification process [131]. The technique is known to be effective if the mask is correctly estimated, but estimating it automatically is hard. Indeed, Wang [132] has proposed that estimation of the time–frequency masks of sound sources can be viewed as the computational goal of auditory scene analysis in general. Fig. 12 illustrates the use of binary masks in the case of a mixture consisting of singing and piano accompaniment. Estimating the mask in music is complicated by the fact that consonant pitch intervals cause partials of different sources to co-incide in frequency. Kitahara *et al.* [133] avoided the mask estimation by applying linear discriminant analysis on features extracted from polyphonic training data. As a result, they obtained feature weightings where the largest weights were given to features that were least affected by the overlapping partials of co-occurring sounds.

A number of systems are based on separating the sound sources from a mixture and then recognizing each of them individually [115], [123], [134]–[136]. Heittola *et al.* [115] used a source-filter model for separating the signals of individual instruments from a mixture. They employed a multiple-F0 estimator to produce candidate F0s at each time instant, and then developed a variant of the non-negative matrix factorization algorithm to assign sounds to their respective instruments

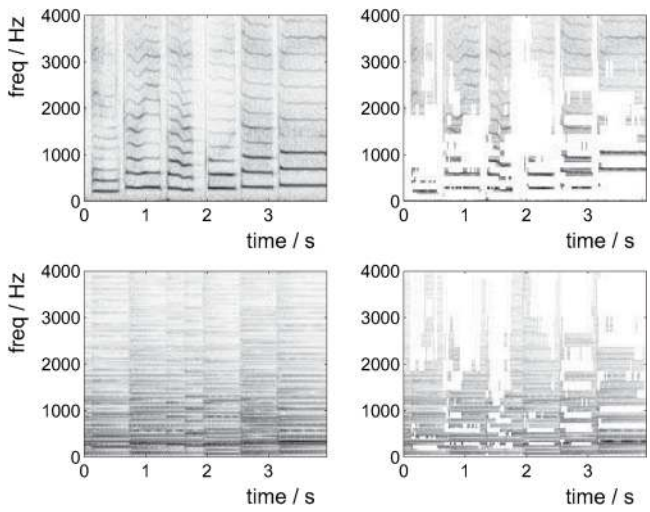

Fig. 12. Illustration of the use of binary masks. The left panels show the magnitude spectrograms of a singing excerpt (top) and its piano accompaniment (bottom). The right panels show spectrograms of the mixture of the singing and the accompaniment, with two different binary masks applied. On the top-right, white areas indicate regions where the accompaniment energy is higher than that of singing. On the bottom-right, singing has been similarly masked out.

and to estimate the spectral envelope of each instrument. A different approach to sound separation was taken by Martins *et al.* [135] and Burred *et al.* [136] who employed ideas from computational auditory scene analysis [137]. They extracted sinusoidal components from the mixture spectrum and then used cues such as common onset time, frequency proximity, and harmonic frequency relationships to assign spectral components to distinct groups. Each group of sinusoidal trajectories was then sent to a recognizer.

Vincent and Rodet [138] viewed instrument recognition as a parameter estimation problem for a given signal model. They represented the short-term log-power spectrum of polyphonic music as a weighted nonlinear combination of typical note spectra plus background noise. The note spectra for each instrument were learnt in advance from a database of isolated notes. Parameter estimation was carried out by maximizing the joint posterior probability of instrument labels and the activation parameters E_{ijt} of note j and instrument i at time t. Maximizing this joint posterior resulted in joint instrument recognition and polyphonic transcription.

A somewhat different path can also be followed by relying on sparse decompositions. The idea of these is to represent a given signal with a small number of elements drawn from a large (typically overcomplete) dictionary. For example, Leveau *et al.* [139] represented a time-domain signal $x(t)$ as a weighted sum of atoms $h_\lambda(t)$ taken from a dictionary $\mathcal{D} = \{h_\lambda(t)\}_\lambda$, and a residual $r(t)$:

$$x(t) = \sum_{\lambda \in \Lambda} a_\lambda h_\lambda(t) + r(t) \tag{11}$$

where Λ is a finite set of indexes λ. Each atom $h_\lambda(t)$ consists of a sum of M windowed and amplitude-weighted sinusoidals at frequencies that are integer multiples of a linearly varying fundamental frequency. An individual atom covers only a short frame of the input signal, but continuity constraints can be placed on the activations a_λ of atoms with successive temporal supports. Leveau *et al.* [139] learned the dictionary of atoms in advance from a database of isolated musical tones. A sparse decomposition for a given mixture signal was then found by maximizing the signal-to-residual ratio for a given number of atoms. This optimization process results in selecting the most suitable atoms from the dictionary, and since the atoms have been labeled with pitch and instrument information, this results in joint instrument identification and polyphonic transcription. Also the instrument recognition methods of Kashino and Murase [140] and Cont and Dubnov [110] can be viewed as being based on dictionaries, the former using time-domain waveform templates and the latter modulation spectra.

The above-discussed sparse decompositions can be viewed as a mid-level representation, where information about the signal content is already visible, but no detection or thresholding has yet taken place. Such a goal was pursued by Kitahara *et al.* [128] who proposed a "note-estimation-free" instrument recognition system for polyphonic music. Their system used a spectrogram-like representation ("instrogram"), where the two dimensions corresponded to time and pitch, and each entry represented the probability that a given target instrument is active at that point.

V. Polyphony and Musical Voices

Given the extensive literature of speech signal analysis, it seems natural that numerous music signal processing studies have focused on monophonic signals. While monophonic signals certainly result in better performance, the desire for wider applicability has led to a gradual focus, in recent years, to the more challenging and more realistic case of polyphonic music. There are two main strategies for dealing with polyphony: the signal can either be processed globally, directly extracting information from the polyphonic signal, or the system can attempt to first split up the signal into individual components (or sources) that can then be individually processed as monophonic signals. The source separation step of this latter strategy, however, is not always explicit and can merely provide a mid-level representation that facilitates the subsequent processing stages. In the following sections, we present some basic material on source separation and then illustrate the different strategies on a selection of specific music signal processing tasks. In particular, we address the tasks of multi-pitch estimation and musical voice extraction including melody, bass, and drum separation.

A. Source Separation

The goal of source separation is to extract all individual sources from a mixed signal. In a musical context, this translates in obtaining the individual track of each instrument (or individual notes for polyphonic instruments such as piano). A number of excellent overviews of source separation principles are available; see [141] and [142].

In general, source separation refers to the extraction of full bandwidth source signals but it is interesting to mention that several polyphonic music processing systems rely on a simplified source separation paradigm. For example, a filter bank decomposition (splitting the signal in adjacent well defined frequency bands) or a mere Harmonic/Noise separation [143] (as for drum extraction [144] or tempo estimation [61]) may be regarded as instances of rudimentary source separation.

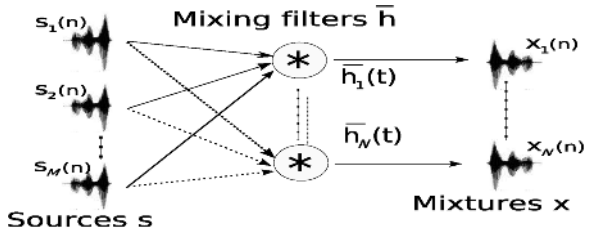

Fig. 13. Convolutive mixing model. Each mixture signal $\mathbf{x}_i(n)$ is then expressed from the source signals as: $\mathbf{x}_i(n) = \sum_{j-1}^{M} h_{ij}(t) * \mathbf{s}_j(n)$.

Three main situations occur in source separation problems. The *determined case* corresponds to the situation where there are as many mixture signals as different sources in the mixtures. Contrary, the *overdetermined* (resp. *underdetermined*) case refers to the situation where there are more (resp. less) mixtures than sources. Underdetermined Source Separation (USS) is obviously the most difficult case. The problem of source separation classically includes two major steps that can be realized jointly: estimating the mixing matrix and estimating the sources. Let $\mathbf{X} = [\mathbf{x}_1(n), \ldots, \mathbf{x}_N(n)]^T$ be the N mixture signals, $\mathbf{S} = [\mathbf{s}_1(n), \ldots, \mathbf{s}_M(n)]^T$ the M source signals, and $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N]^T$ the $N \times M$ mixing matrix with mixing gains $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{iM})$. The mixture signals are then obtained by: $\mathbf{X} = \mathbf{AS}$. This readily corresponds to the *instantaneous mixing* model (the mixing coefficients are simple scalars). The more general convolutive mixing model considers that a filtering occurred between each source and each mixture (see Fig. 13). In this case, if the filters are represented as $N \times M$ FIR filters of impulse response $h_{ij}(t)$, the mixing matrix is given by $\mathbf{A} = (\overline{\mathbf{h}}_1(t), \overline{\mathbf{h}}_2(t), \ldots, \overline{\mathbf{h}}_N(t))^T$ with $\overline{\mathbf{h}}_i(t) = [h_{i1}(t), \ldots, h_{iM}(t)]$, and the mixing model corresponds to $\mathbf{X} = \mathbf{A} * \mathbf{S}$.

A wide variety of approaches exist to estimate the mixing matrix and rely on techniques such as Independent Component Analysis (ICA), sparse decompositions or clustering approaches [141]. In the determined case, it is straightforward to obtain the individual sources once the mixing matrix is known: $\mathbf{S} = \mathbf{A}^{-1}\mathbf{X}$. The underdetermined case is much harder since it is an ill-posed problem with an infinite number of solutions. Again, a large variety of strategies exists to recover the sources including heuristic methods, minimization criteria on the error $||\mathbf{X} - \mathbf{AS}||^2$, or time–frequency masking approaches. One of the popular approaches, termed adaptive Wiener filtering, exploits soft time–frequency masking. Because of its importance for audio source separation, it is described in more details.

For the sake of clarity, we consider below the monophonic case, i.e., where only one mixture signal $\mathbf{x}_1(n)$ is available. If we consider that the M sources $\mathbf{s}_i(n)$ are stationary Gaussian processes of power spectral density (PSD) $\sigma_i^2(k)$, then the optimal estimate of $\mathbf{s}_i(n)$ is obtained as

$$S_i(n,k) = \frac{\sigma_i^2(k)}{\sum_{j=1}^{M} \sigma_j^2(k)} X_1(n,k) \qquad (12)$$

where $X_1(n,k)$ and $S_i(n,k)$ are the STFTs of the mixture $\mathbf{x}_1(n)$ and source $\mathbf{s}_i(n)$, respectively. In practice, audio signals can only be considered as locally stationary and are generally assumed to be a combination of stationary Gaussian processes. The source signal $\mathbf{s}_i(n)$ is then given by

$$\mathbf{s}_i(n) = \sum_{m \in K_i} \alpha_m(n) b_m(n),$$

where $b_m(n)$ are stationary Gaussian processes of PSD $\sigma_m^2(k)$, $\alpha_m(n) \geq 0$ are slowly varying coefficients, and K_i is a set of indices for source $\mathbf{s}_i(n)$. Here, the estimate of $\mathbf{s}_i(n)$ is then obtained as (see for example [145] or [144] for more details):

$$S_i(n,k) = \frac{\sum_{m \in K_i} \alpha_m(n) \sigma_m^2(k)}{\sum_{m \in K_1 \cup \ldots \cup K_M} \alpha_m(n) \sigma_m^2(k)} X_1(n,k). \qquad (13)$$

Note that in this case, it is possible to use decomposition methods on the mixture $\mathbf{x_1}(n)$ such as non-negative matrix factorization (NMF) to obtain estimates of the spectral templates $\sigma_m^2(k)$.

Music signal separation is a particularly difficult example of USS of convolutive mixtures (many concurrent instruments, possibly mixed down with different reverberation settings, many simultaneous musical notes and, in general, a recording limited to two channels). The problem is then often tackled by integrating prior information on the different source signals. For music signals, different kinds of prior information have been used including timbre models [146], harmonicity of the sources [147], temporal continuity, and sparsity constraints [148]. In some cases, by analogy with speech signal separation, it is possible to exploit production models, see [114] or [149].

Concerning evaluation, the domain of source separation of audio signals is also now quite mature and regular evaluation campaigns exist[5] along with widely used evaluation protocols [150].

B. From Monopitch to Multipitch Estimation

The estimation of the fundamental frequency of a quasi-periodic signal, termed *monopitch estimation*, has interested the research community for decades. One of the main challenges is to obtain a versatile and efficient algorithm for a wide range of possible fundamental frequencies which can cope with the deviations of real audio signals from perfect periodicity. For speech signals, extensive reviews of early algorithms can be found in [151] and [152].

In general, the different methods can be roughly classified in three classes depending on the used signal representation. The *frequency domain approaches* exploit the fact that quasi-periodic signals exhibit a quasi-harmonic distribution of peaks in the spectral domain. The fundamental frequency is estimated by searching the highest frequency that generates a spectral comb best explaining the spectral content of the signal; see [153]–[155]. The *time domain approaches* aim at directly estimating the period on the basis of the signal's waveform by searching the smallest time-shift for which the waveform and its time-shifted version match. This can be done using the autocorrelation or the average magnitude difference functions [156], [157] or applying kernel-based approaches [158]. Both time and frequency domain approaches are prone to octave

[5]For example, see http://sisec.wiki.irisa.fr/tiki-index.php (SiSec campaign).

errors. However, because frequency domain (resp. time-domain) approaches are prone to estimate integer multiples or harmonics (resp. integer fractions or sub-harmonics) of the true fundamental frequency, *mixed domain approaches* exploiting both representations were also developed; see [159] or [160].

Even though monopitch estimation is now achievable with a reasonably high accuracy, the problem of multipitch estimation (e.g., estimating the fundamental frequency of concurrent periodic sounds) remains very challenging. The problem is indeed particularly difficult for music signals for which concurrent notes stand in close harmonic relation. Here, in some sense, the worst case is when two simultaneous notes are played one or several octaves apart. For extreme cases such as complex orchestral music, where one has a high level of polyphony, multipitch estimation becomes intractable with today's methods. For a review of recent approaches, we refer to [161]–[163]. Most approaches work, at least partially, in the spectral domain. On the one hand, some methods follow a global strategy aiming at jointly estimating all fundamental frequencies. For example, [164] or [165] employ parametric methods, [166] describes a dedicated comb approach, [167] and [168] use methods based on machine learning paradigms, whereas [169] follows a least-square strategy. On the other hand, a number of methods rely on source separation principles which are more or less explicit. For example, the non-negative matrix factorization framework was successfully used in several algorithms [147], [148], [170]. In other approaches, the source separation may be less explicit and can, for example, rely on an iterative procedure [171]. Here, the dominant fundamental frequency is first estimated, then the spectral peaks of the corresponding musical note are identified and subtracted (sometimes only partially) from the polyphonic signal to obtain a residual signal. The procedure is iterated while the residual contains at least one musical note. Despite the inherent limitations of iterative procedures, these approaches are among the most efficient to date as regularly shown in the MIREX evaluation campaign.[6]

It is worth emphasizing that most methods exploit musical knowledge in one or the other way. Characteristic timbre information may be used in the form of instrument spectral models or templates as prior information to better separate the musical sources (as in [172]). Also, spectral smoothness principles can be exploited to subtract more realistic musical notes in iterative methods (as in [171]). Furthermore, constraints on the synchronous evolution of partials amplitude may further help to identify the spectral contribution of a note in a mixture [165]. The specificities of the production mechanism can also be exploited, e.g., in the form of a source/filter production model for the separation, or the introduction of inharmonicities in the model for instruments such as piano [173]. It is also possible to reach higher performances by means of duration or note evolution models (for example based on hidden Markov models as in [173], [174]). This, indeed, permits to take advantage of observations in successive time frames. Finally, knowledge of auditory perception has also been used with success in a number of methods mostly as a front-end acoustic analysis, see [159] and [175]. It is believed that future progress will be fueled by a better understanding of the perception of sound mixtures. Here, an interesting question is why a trained musician has no problem in analyzing a chord containing two notes one octave apart. Better understanding in which way two overlapping partials interact and how their amplitude can be precisely estimated is certainly of central importance [176].

[6]See http://www.music-ir.org/mirex/wiki/.

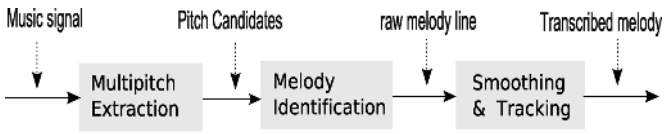

Fig. 14. General scheme underlying most melody transcription systems. From [183].

C. Main Melody or Singing Voice Extraction

Main melody extraction, especially for the singing voice, has received a great deal of interest. This is mainly motivated by the wide range of potential applications including karaoke [177], query-by-humming [178], lead-sheet generation [179], query-by-examples and cover version detection [12], [33], [180]. Following the definition by Paiva [181], "*Melody is the dominant individual pitched line in a musical ensemble*," the task is often restricted to a mere predominant-F0 estimation and tracking task [182]. Only few studies address the full problem leading to a musical score of the melody line, which integrates a note segmentation stage [174], [179]. As described in [183], the problem of main melody extraction is traditionally split into a preliminary analysis stage followed by a melody identification phase and concluded by a smoothing or tracking process; see Fig. 14.

The analysis stage can directly output a raw sequence of predominant fundamental frequency candidates (as in [174] or [181]) but can also produces an intermediate representation or probabilistic model with posterior probabilities for each potential note that would be further exploited in the melody tracking stage (as in [182] or [114]). The analysis stage mostly relies on a spectral domain representation as obtained by a traditional short-time Fourier transform, but may also be obtained by specific multi-resolution transforms [184] or perceptually motivated representations [181].

In the melody identification phase, the sequence of fundamental frequencies that most likely corresponds to the melody is identified using, for example, ad hoc rules, constrained dynamic programming, or hidden Markov models. Finally, if not integrated in the previous stage, a final smoothing or tracking process occurs where the initial estimation of the melody line is further smoothed in order to avoid sudden jumps in the melody line. Such jumps may be caused by initial octave errors or other extraction errors.

In terms of performance, it appears that the most accurate algorithm evaluated in the MIREX evaluation campaigns follows a rule based approach [185] although statistical based systems indicate very promising directions for the future.[7]

To provide some insight on main melody extraction, we now describe one of the existing statistical approaches in more detail [114], [186]. For the sake of clarity, we consider here the

[7]www.music-ir.org/mirex/wiki/2009:Audio_Melody_Extraction_Results.

monophonic case, where a single mixture signal $\mathbf{x}_1(n)$ is observed. In this model, the observed signal $\mathbf{x}_1(n)$ is the sum of two contributions: the leading voice $\mathbf{v}(n)$ and the musical background $\mathbf{m}(n)$. For a given frame, the STFT of the mixture signal $X(n,k)$ can then be expressed as $X(n,k) = V(n,k) + M(n,k)$, where $V(n,k)$ and $M(n,k)$ are the STFTs of $\mathbf{v}(n)$ and $\mathbf{m}(n)$, respectively. Furthermore, $V(n,k)$ and $M(n,k)$ are assumed to be center proper Gaussians:[8]

$$V(n,k) \sim \mathcal{N}_c\left(0, \mathrm{diag}\left(\sigma^2_{n,V}(k)\right)\right) \quad (14)$$

$$M(n,k) \sim \mathcal{N}_c\left(0, \mathrm{diag}\left(\sigma^2_{n,M}(k)\right)\right) \quad (15)$$

where $\sigma^2_{n,V}(k)$ (resp. $\sigma^2_{n,M}(k)$) is the power spectral density (PSD) of the leading voice $\mathbf{v}(n)$ (resp. of the background music $\mathbf{m}(n)$). Assuming that the leading voice and the background music are independent, the mixture signal at frame n is also a proper Gaussian vector:

$$X(n,k) \sim \mathcal{N}_c\left(0, \mathrm{diag}\left(\sigma^2_{n,X}(k)\right)\right) \quad (16)$$

$$\sim \mathcal{N}_c\left(0, \mathrm{diag}\left(\sigma^2_{n,V}(k) + \sigma^2_{n,M}(k)\right)\right). \quad (17)$$

For extracting the main melody, one then needs to estimate $\sigma^2_{n,V}(k)$ and $\sigma^2_{n,M}(k)$. This can be done by expressing the PSDs as Gaussian mixture models learned on dedicated databases [145]. In [114], assuming a singing voice, the approach is entirely unsupervised, i.e., no learning step is involved. Instead it relies on specific constraints for the voice and the musical background signal. More precisely, the voice signal is assumed to follow a source/filter production model where the source is a periodic signal (referring to the periodic glottal pulse of the singing voice) and where the filter is constrained to smoothly evolve (referring to the slowly varying vocal tract shapes while singing). For the musical background signal, no specific constraints are assumed because of the wide variability of possible music instruments. The estimation of the various model parameters is then conducted by iterative approaches based on NMF techniques. Once the PSDs $\sigma^2_{n,V}(k)$ and $\sigma^2_{n,M}(k)$ of both signals are obtained, the separated singing voice signal is obtained using the Wiener filter approach for each frame as in (12).

Since the model is rather generic, it is also directly applicable to other leading instruments such as a trumpet within a Jazz quartet; see Fig. 15.

D. Bass Line Extraction

Following [174], the term *bass line* refers to an organized sequence of consecutive notes and rests played with a bass guitar, a double bass or a bass synthesizer. The bass line plays an important role in several music styles particularly in popular music. Having a separated bass line or a transcription of the bass line opens the path to a number of applications including "music minus one" for bass players or various indexing tasks such as chord extraction, downbeat estimation, music genre or mood classification [187].

[8]A complex proper Gaussian random variable is a complex random variable whose real part and imaginary part are independent and follow a (real) Gaussian distribution, with the same parameters: mean equal to 0 and identical variance (co-variance matrix in the multi-variate case).

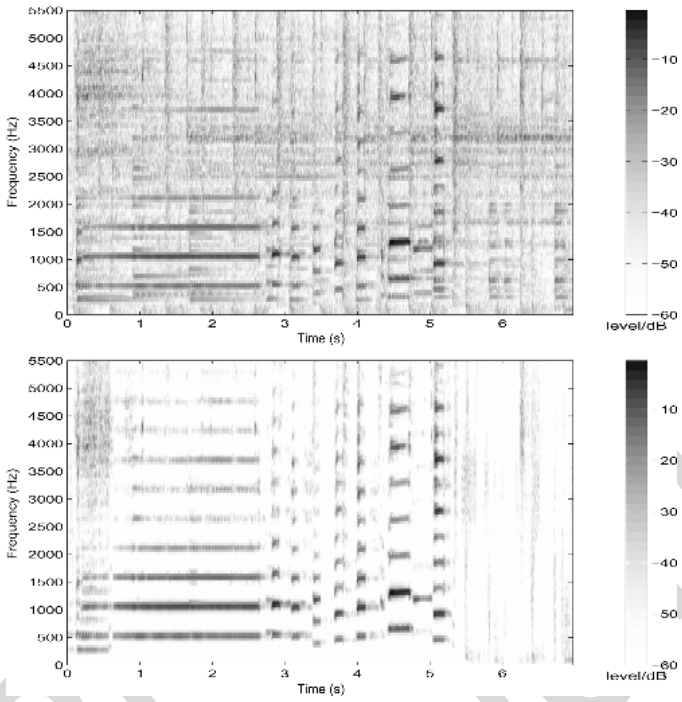

Fig. 15. Spectrograms of the original signal (top) and of the separated trumpet signal (bottom) of the piece *Caravan* played by the *Marsalis Jazz Quartet*. From [114].

Bass line transcription is amongst the earliest studies on transcribing a single instrument track from a rich, polyphonic music signal [188], [189]. The problem of bass line transcription, which typically refers to the lower frequency range between 30 and 250 Hz, bears many similarities with the main melody extraction problem. Indeed, as for melody, the task is often regarded as a mere predominant-F0 estimation, where the tracking is now done in the lower frequency range. Not surprisingly, there are a number of approaches that were proposed to extract both melody and bass line within the same general framework [17], [174].

As an example, we now describe the system proposed in [190] in more detail. It is based on two frame-based feature extractors: namely a multiple-F0 estimator that provides salience value for four F0 estimates and an accent estimator that measures the probability of having an onset. One of the specificities of this approach is the integration of two different models: a note and a rest model. Bass notes are modeled using a three-state left-right HMM (the three states aim at capturing the attack, sustain and release phases of a played note) while rest notes are represented by a four-component GMM model (equivalent to a single state HMM). In subsequent work, a third background model was included to better represent the notes played by other instruments [174]. Another specificity of this model is the use of a musicological model that controls transition probabilities between the note models and the rest model. These transition probabilities depend on the musical key and on the preceding notes. The musical key is estimated from the set of four F0 candidates over the entire history of the piece in combination with simple artificial key profiles. The sequence model is either a bi-gram model or a more sophisticated variable-order Markov Model learned

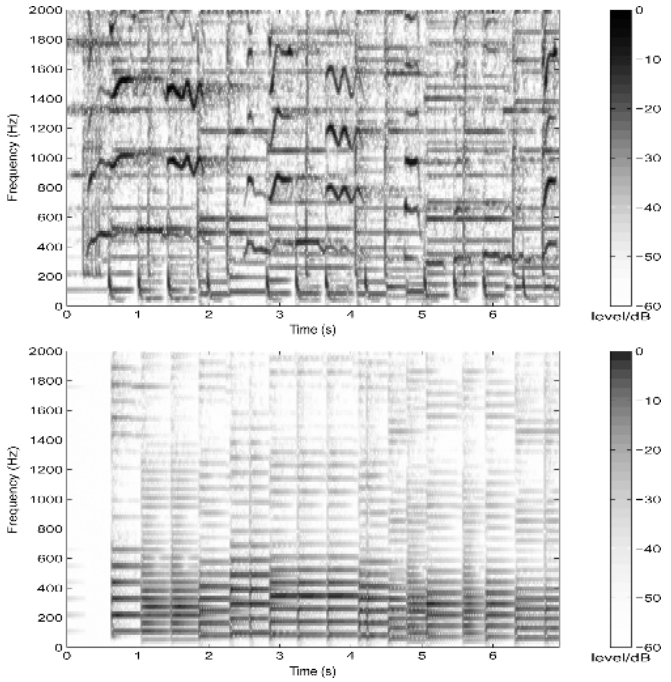

Fig. 16. Spectrograms of the original signal (top) and of the bass-line (bottom) resynthesized from its estimated MIDI transcription. From [190].

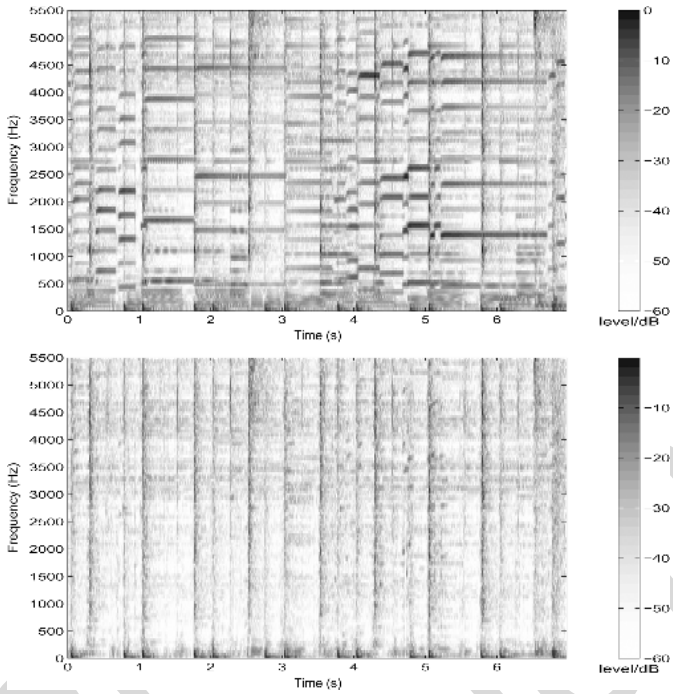

Fig. 17. Spectrograms of the original signal (top) and of the separated drum signal (bottom) obtained by enhanced Wiener filtering. From [144].

on a collection of MIDI files. Finally, Viterbi decoding is performed to obtain the most probable path through the models which yields a transcription of the bass line. As an illustration, Fig. 16 shows the spectrograms of a polyphonic music signal and of the bass-line resynthesized from its estimated MIDI transcription.

E. Drum Extraction

Historically, the analysis of the percussive or drum component of music signals has attracted less attention from the research community. Nevertheless, it has now been recognized that numerous applications can benefit from an appropriate processing of this percussive component including beat tracking, content-based retrieval based on the query-by-tapping paradigm, or beatboxing [191]–[194]. Furthermore, the rhythmic content, if successfully extracted, is crucial for developing music similarity measures as needed in music classification tasks. Even though some studies have focussed on traditional percussive instruments (see for example [195]–[197] for Indian percussions), most research has targeted the western drum kit composed of at least three main instrument classes (bass drum, snare drum, and cymbals). Starting from solo drum signals (see [198] for a review), most of the recent studies tackle the more realistic scenario of extracting and transcribing drum signals directly from polyphonic signals.

Following [144], the different approaches can be classified in three categories: *segment and classify*, *match and adapt*, or *separate and detect*. The *segment and classify* approaches either first segment the signal into individual discrete segments which are then classified using machine learning techniques [199]–[201] or jointly perform the two steps using, for example, hidden Markov models [202]. The *match and adapt* approaches rather aim at searching for occurrences of reference templates in the music signal which can be further adapted to the specificity of the analyzed signal [203], [204]. Finally, the *separate and detect* approaches first aim at separating or extracting the percussive component before further analysis to identify the drum events. As it has turned out, the approaches based on Independent Subspace Analysis (ISA) [205], [206], or [207] or non-negative matrix factorization [208], [209] are among the most successful systems for drum track separation.

To illustrate some of the above ideas, we now describe the drum track separation approach proposed in [144] in more detail. This algorithm is based on the general Wiener filtering approach for source separation, optimized for drum separation. The separation itself is performed using (13) but exploits different strategies for learning the spectral templates $\sigma_m^2(f)$ of each source. The drum spectral templates are learned on solo drum signals by non-negative matrix factorization (NMF) while the spectral templates for the background music are learned by a correlation-based clustering algorithm (as in [210]). A total of 144 templates are then learned—128 for the background music and 16 for the drum component. Then, an adaptation procedure is applied to better cope with the inherent variability of real audio signals. This adaptation consists in extending the set of 16 learned drum spectral templates by the PSD of the stochastic component obtained by subband subspace projection [211]. Indeed, this additional template already provides a decent estimate of the PSD of the drum signal and therefore facilitates the convergence of the algorithm to an appropriate solution. Finally, in order to represent at the same time the short drum onsets and steady part of tonal components, a multi-resolution approach is followed by implementing a window-size switching scheme for time–frequency decomposition based on the output of a note onset detection algorithm.

Fig. 17 gives an example of the result of the above algorithm by displaying the spectrograms of a polyphonic music signal and its separated drum signal.

VI. Conclusion

Signal processing for music analysis is a vibrant and rapidly evolving field of research, which can enrich the wider signal processing community with exciting applications and new problems. Music is arguably the most intricate and carefully constructed of any sound signal, and extracting information of relevance to listeners therefore requires the kinds of specialized methods that we have presented, able to take account of music-specific characteristics including pitches, harmony, rhythm, and instrumentation.

It is clear, however, that these techniques are not the end of the story for analyzing music audio, and many open questions and research areas remain to be more fully explored. Some of the most pressing and promising are listed below.

- Decomposing a complex music signal into different components can be a powerful preprocessing step for many applications. For example, since a cover song may preserve the original melody but alter the harmonization, or alternatively keep the same basic chord progression but devise a new melody (but rarely both), a promising approach to cover version recognition is to separate the main melody and accompaniment, and search in both parts independently [180]. A different decomposition, into harmonic and percussive components, has brought benefits to tasks such as chord recognition [22], genre classification [105], and beat tracking [61]. Note that such decompositions need not be perfect to yield benefits—even a modest improvement in the relative level between components can give a significant improvement in a subsequent analysis.
- Improved recognition and separation of sources in polyphonic audio remains a considerable challenge, with great potential to improve both music processing and much broader applications in computational auditory scene analysis and noise-robust speech recognition. In particular, the development, adaptation, and exploitation of sound source models for the purpose of source separation seems to be required in order to achieve an accuracy comparable to that of human listeners in dealing with polyphonic audio [212].
- In conjunction with the appropriate signal processing and representations, machine learning has had some great successes in music signal analysis. However, many areas are limited by the availability of high-quality labeled data. In chord recognition, for instance, the entire field is using the same corpus of 200 tracks for which high-quality manual chord transcripts have been prepared [23]. However, while special-purpose human labeling remains the gold standard, it is interesting to note that a given piece of music may have multiple, closely related sources of information, including alternate recordings or performances, partial mixes derived from the original studio multitracks, score representations including MIDI versions, lyric transcriptions, etc. These different kinds of information, some available in large quantities, present opportunities for innovative processing that can solve otherwise intractable problems such as score-guided separation [213], generate substitutes for manual ground-truth labels using music synchronization techniques [28], [29], [32], or use multi-perspective approaches to automatically evaluate algorithms [82], [214].
- Source separation and audio transcription, despite their obvious relationship, are often tackled as independent and separate tasks: As we have shown, a number of music signal analysis systems include some level of source separation. Other work in source separation has shown that performance is usually improved when appropriate prior information is used—information such as musical scores. Rather than relying on existing, ground-truth scores, information of this kind could also be obtained from rudimentary (or more elaborate) automatic transcription. Significant progress can perhaps be made in both fields by better exploiting transcription in source separation (so-called "informed" source separation) and by better integrating source separation in transcription systems.
- Many music analysis tasks have encountered a "glass ceiling," a point beyond which it has become very difficult to make improvements. One tactic is to restrict the domain, to allow an approach to specialize on a limited subset—for example, by building a beat tracker that is specialized for jazz, and a different one for classical music. This suggests a broader strategy of deploying a context-adaptation layer, able to choose parameters and models best suited to each particular signal. In the simplest case, this can be implemented by training the methods separately for, say, different genres, and then using automatic audio classification to choose the best models for a given test signal, but how to implement a more general and optimal context adaptation is a deep and open research question.
- Surprisingly, knowledge about auditory perception has a limited role is most music signal processing systems, but since music exists purely to be heard, hearing science promises to advance our understanding music perception and should therefore inform the analysis of complex signals such as polyphonic mixtures. In multipitch estimation, for example, understanding the way that overlapping partials interact and how their amplitudes can be precisely estimated represents one promising direction [176], [215].
- Much current research focuses on individual aspects of the music (e.g., the rhythm, or the chords, or the instruments). These aspects, however, are anything but independent, and we expect significant synergies from efforts to analyze them jointly, with information from one aspect helping to improve the extraction of another. Some examples of this include approaches that jointly use metric, harmonic, and structural cues to support and stabilizing tempo and beat tracking [20], [54], [75], [216], [217].

The work described in this paper illustrates the broad range of sophisticated techniques that have been developed in the rapidly evolving field of music signal analysis, but as shown by this list of open questions, there is much room for improvement and for new inventions and discoveries, leading to more powerful and innovative applications. While, for the moment, human listeners

remain far superior to machines in extracting and understanding the information in music signals, we hope that continued development of automatic techniques will lessen this gap, and may even help to clarify some aspects of how and why people listen to music.

ACKNOWLEDGMENT

[AUTHOR: Please provide end page]

REFERENCES

[1] C. Roads, *The Computer Music Tutorial*. Cambridge, MA: MIT Press, 1996.

[2] M. S. Puckette, *The Theory and Technique of Electronic Music*. Singapore: World Scientific, 2007.

[3] J. A. Moorer, "On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer," Ph.D. dissertation, Department of Music, Stanford Univ., Stanford, CA, 1975.

[4] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. New York: Academic, 2003.

[5] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*. New York: Springer-Verlag, 1990.

[6] M. Müller, *Information Retrieval for Music and Motion*. New York: Springer-Verlag, 2007.

[7] K. L. Kashima and B. Mont-Reynaud, "The bounded-Q approach to time-varying spectral analysis," Dept. of Music, Stanford Univ., Tech. Rep. STAN-M-28, 1985.

[8] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Amer. (JASA)*, vol. 92, pp. 2698–2701, 1992.

[9] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. Sound Music Comput. Conf. (SMC)*, Barcelona, Pain, 2010.

[10] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proc. ICMC*, Beijing, China, 1999, pp. 464–467.

[11] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, 2001, pp. 15–18.

[12] D. P. W. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Honolulu, HI, Apr. 2007, vol. 4, pp. 1429–1439.

[13] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 135–140.

[14] M. Müller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 649–662, Mar. 2010.

[15] A. Sheh and D. P. W. Ellis, "Chord segmentation and recognition using EM-trained hidden Markov models," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Baltimore, MD, 2003.

[16] E. Gómez, "Tonal description of music audio signals" Ph.D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2006 [Online]. Available: files/publications/emilia-PhD-2006.pdf

[17] 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. (ISCA J.)*, vol. 43, no. 4, pp. 311–329, 2004.

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

[19] K. Lee and M. Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 291–301, Feb. 2008.

[20] H. Papadopoulos and G. Peeters, "Simultaneous estimation of chord progression and downbeats from an audio file," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 121–124.

[21] M. Mauch, K. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Kobe, Japan, 2009, pp. 231–236.

[22] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama, "HMM-based approach for automatic chord detection using refined acoustic features," in *Proc. 35nd IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, 2010, pp. 5518–5521.

[23] C. Harte, M. Sandler, S. Abdallah, and E. Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005.

[24] S. Ewert, M. Müller, and R. B. Dannenberg, "Towards reliable partial music alignments using multiple synchronization strategies," in *Proc. Int. Workshop Adaptive Multimedia Retrieval (AMR)*, Madrid, Spain, Sep. 2009.

[25] D. Damm, C. Fremerey, F. Kurth, M. Müller, and M. Clausen, "Multimodal presentation and browsing of music," in *Proc. 10th Int. Conf. Multimodal Interfaces (ICMI)*, Chania, Crete, Greece, Oct. 2008, pp. 205–208.

[26] R. Dannenberg, "An on-line algorithm for real-time accompaniment," in *Proc. Int. Comput. Music Conf. (ICMC)*, 1984, pp. 193–198.

[27] M. Müller, H. Mattes, and F. Kurth, "An efficient multiscale approach to audio synchronization," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR)*, Victoria, BC, Canada, Oct. 2006, pp. 192–197.

[28] R. J. Turetsky and D. P. Ellis, "Ground-truth transcriptions of real music from force-aligned MIDI syntheses," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Baltimore, MD, 2003, pp. 135–141.

[29] N. Hu, R. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, Oct. 2003.

[30] S. Dixon and G. Widmer, "Match: A music alignment tool chest," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005.

[31] C. Fremerey, F. Kurth, M. Müller, and M. Clausen, "A demonstration of the SyncPlayer system," in *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR)*, Vienna, Austria, Sep. 2007, pp. 131–132.

[32] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 1869–1872.

[33] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1138–1151, Aug. 2008.

[34] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, pp. 195–197, 1981.

[35] M. Casey and M. Slaney, "Fast recognition of remixed music audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, pp. 1425–1428.

[36] M. Casey, C. Rhodes, and M. Slaney, "Analysis of minimum distances in high-dimensional musical spaces," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 1015–1028, Jul. 2008.

[37] F. Kurth and M. Müller, "Efficient index-based audio matching," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 382–395, Feb. 2008.

[38] D. B. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press, 2006.

[39] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *Proc. 11th Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 625–636.

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

[41] M. Müller and F. Kurth, "Towards structural analysis of audio recordings in the presence of musical variations," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, p. 163, 2007.

[42] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 318–326, Feb. 2008.

[43] J. Paulus and A. Klapuri, "Music structure analysis using a probabilistic fitness measure and a greedy search algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1159–1170, Aug. 2009.

[44] R. J. Weiss and J. P. Bello, "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 123–128.

[45] T. Bertin-Mahieux, R. J. Weiss, and D. P. W. Ellis, "Clustering beat-chroma patterns in a large music database," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 111–116.

[46] R. Parncutt, "A perceptual model of pulse salience and metrical accent in musical rhythms," *Music Percept.*, vol. 11, pp. 409–464, 1994.

[47] W. A. Sethares, *Rhythm and Transforms*. New York: Springer, 2007.

[48] F. Lerdahl and R. Jackendoff, *Generative Theory of Tonal Music*. Cambridge, MA: MIT Press, 1983.

[49] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[50] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *Proc. AES Conv. 118*, Barcelona, Spain, 2005.

[51] R. Zhou, M. Mattavelli, and G. Zoia, "Music onset detection based on resonator time frequency image," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1685–1695, Nov. 2008.

[52] N. Collins, "Using a pitch detector for onset detection," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005, pp. 100–106.

[53] A. J. Eronen and A. P. Klapuri, "Music tempo estimation with k-NN regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 50–57, Jan. 2010.

[54] 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.

[55] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Sep. 2008.

[56] C. C. Toh, B. Zhang, and Y. Wang, "Multiple-feature fusion based onset detection for solo singing voice," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, 2009.

[57] A. P. Klapuri, A. J. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 342–355, Jan. 2006.

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

[59] P. Masri and A. Bateman, "Improved modeling of attack transients in music analysis-resynthesis," in *Proc. Int. Comput. Music Conf. (ICMC)*, Hong Kong, 1996, pp. 100–103.

[60] M. Alonso, B. David, and G. Richard, "Tempo and beat estimation of musical signals," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[61] M. Alonso, G. Richard, and B. David, "Accurate tempo estimation based on harmonic+noise decomposition," *EURASIP J. Adv. Signal Process.*, vol. 2007, p. 14, 2007, Article ID 82 795.

[62] P. Grosche and M. Müller, "A mid-level representation for capturing dominant tempo and pulse information in music recordings," in *Proc. 10th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, Oct. 2009, pp. 189–194.

[63] P. Grosche and M. Müller, "Extracting predominant local pulse information from music recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, 2011, to be published.

[64] 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. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1832–1844, Sep. 2006.

[65] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1009–1020, Mar. 2007.

[66] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, 2007.

[67] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, p. 158, 2007.

[68] J. Seppänen, A. Eronen, and J. Hiipakka, "Joint beat & tatum tracking from music signals," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Victoria, Canada, 2006.

[69] ***[AUTHOR: Please provide end page]***A. T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempogram representation and Kalman filtering," *J. New Music Res.*, vol. 28, no. 4, pp. 259–273, 2001.

[70] K. Jensen, J. Xu, and M. Zachariasen, "Rhythm-based segmentation of popular chinese music," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005.

[71] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *Proc. Int. Conf. Multimedia Expo (ICME)*, Los Alamitos, CA, 2001.

[72] F. Kurth, T. Gehrmann, and M. Müller, "The cyclic beat spectrum: Tempo-related audio features for time-scale invariant audio identification," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR)*, Victoria, BC, Canada, Oct. 2006, pp. 35–40.

[73] P. Grosche, M. Müller, and F. Kurth, "Cyclic tempogram—A mid-level tempo representation for music signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, Mar. 2010, pp. 5522–5525.

[74] G. Peeters, "Time variable tempo detection and beat marking," in *Proc. Int. Comput. Music Conf. (ICMC)*, Barcelona, Spain, 2005.

[75] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *J. New Music Res.*, vol. 30, pp. 39–58, 2001.

[76] J. Seppänen, "Tatum grid analysis of musical signals," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2001, pp. 131–134.

[77] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Paris, France, 2002, pp. 150–156.

[78] N. Degara, A. Pena, M. E. P. Davies, and M. D. Plumbley, "Note onset detection using rhythmic structure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, 2010, pp. 5526–5529.

[79] J. Bilmes, "Techniques to foster drum machine expressivity," in *Proc. Int. Comput. Music Conf.*, Tokyo, Japan, 1993.

[80] F. Gouyon and P. Herrera, "Pulse-dependent analysis of percussive music," in *Proc. AES 22nd Int. Conf. Virtual, Synthetic Entertainment Audio*, Espoo, Finland, 2002.

[81] S. Dixon and W. Goebl, "Pinpointing the beat: Tapping to expressive performances," in *Proc. Int. Conf. Music Percep. Cogn.*, Sydney, Australia, 2002, pp. 617–620.

[82] P. Grosche, M. Müller, and C. S. Sapp, "What makes beat tracking difficult? A case study on chopin mazurkas," in *Proc. 11th Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 649–654.

[83] E. W. Large and C. Palmer, "Perceiving temporal regularity in music," *Cognitive Sci.*, vol. 26, no. 1, pp. 1–37, 2002.

[84] S. Dixon, F. Gouyon, and G. Widmer, "Towards characterization of music via rhythmic patterns," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[85] G. Peeters, "Rhythm classification using spectral rhythm patterns," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005, pp. 644–647.

[86] F. Gouyon and P. Herrera, "Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *AES Conv. 114*, Amsterdam, The Netherlands, 2003.

[87] S. Dixon, "Evaluation of the audio beat tracking system beatroot," *J. New Music Res.*, vol. 36, pp. 39–50, 2007.

[88] USA Standard Acoustical Terminology American National Standards Inst., Tech. Rep. S1.1-1960, 1960.

[89] J.-J. Aucouturier, "Dix expériences sur la modélisation du timbre polyphonique [Ten Experiments on the Modelling of Polyphonic Timbre]," Ph.D. dissertation, Univ. Paris 6, Paris, France, 2006.

[90] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Percept.*, vol. 27, no. 3, pp. 223–241, 2010.

[91] R. Erickson, *Sound Structure in Music*. Berkeley, CA: Univ. of California, 1975.

[92] J. Barbour, "Analytic listening: A case study of radio production," in *Proc. Int. Conf. Auditory Display*, Sydney, Australia, Jul. 2004.

[93] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 133–141, Mar. 2006.

[94] Y. E. Kim, E. M. Schmidt, R. Migneco, B. C. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 255–266.

[95] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 467–476, Mar. 2008.

[96] J. F. Schouten, "The perception of timbre," in *Proc. 6th Int. Congr. Acoust.*, Tokyo, Japan, 1968, p. GP-6-2.

[97] S. Handel, "Timbre perception and auditory object identification," in *Hearing—Handbook of Perception and Cognition*, B. C. J. Moore, Ed., 2nd ed. San Diego, CA: Academic, 1995, pp. 425–460.

[98] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," *J. Acoust. Soc. Amer.*, vol. 118, no. 1, pp. 471–482, 2005.

[99] R. Cogan, *New Images of Musical Sound*. Cambridge, MA: Harvard Univ. Press, 1984.

[100] R. A. Kendall and E. C. Carterette, "Perceptual scaling of simultaneous wind instrument timbres," *Music Percept.*, vol. 8, no. 4, pp. 369–404, 1991.
[101] A. Eronen, "Comparison of features for musical instrument recognition," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2001, pp. 19–22.
[102] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, ser. Signal Processing Series. Englewood Cliffs: Prentice-Hall, 1993.
[103] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. D. Poli, Eds. Lisse, The Netherlands: Swets & Zeitlinger, 1997.
[104] N. Ono, K. Miyamoto, J. LeRoux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. Eur. Signal Process. Conf.*, Lausanne, Switzerland, 2008, pp. 240–244.
[105] H. Rump, S. Miyabe, E. Tsunoo, N. Ono, and S. Sagama, "Autoregressive MFCC models for genre classification improved by harmonic-percussion separation," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 87–92.
[106] *Hearing—Handbook of Perception and Cognition*, B. C. J. Moore, Ed., 2nd ed. San Diego, CA: Academic, 1995.
[107] M. S. Vinton and L. E. Atlas, "Scalable and progressive audio codec," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, 2001, pp. 3277–3280.
[108] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 576–588, Mar. 2010.
[109] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, 1997, pp. 209–212.
[110] A. Cont, S. Dubnov, and D. Wessel, "Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negativity constraints," in *Proc. 10th Int. Conf. Digital Audio Effects*, Bordeaux, France, 2007, pp. 85–92.
[111] V. Välimäki, J. Pakarinen, C. Erkut, and M. Karjalainen, "Discrete-time modelling of musical instruments," *Rep. Progr. Phys.*, vol. 69, no. 1, pp. 1–78, 2006.
[112] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Tampa, FL, 1985, pp. 937–940.
[113] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, Honolulu, HI, 2007, pp. 53–56.
[114] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 564–575, Mar. 2010.
[115] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. 10th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, 2009, pp. 327–332.
[116] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale*. London, U.K.: Springer, 1998.
[117] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
[118] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[119] P. Herrera-Boyer, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2006, pp. 163–200.
[120] T. Kitahara, "Computational musical instrument recognition and its application to content-based music information retrieval," Ph.D. dissertation, Kyoto Univ., , Japan, Mar. 2007.
[121] A. Eronen, "Signal processing method for audio classification and music content analysis," Ph.D. dissertation, Tampere Univ. of Technol., Tampere, Finland, Jun. 2009.
[122] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 174–186, 2009.
[123] A. A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Proc. Int. Conf. Digital Audio Effects*, Naples, Italy, 2004.
[124] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 68–80, 2006.
[125] G. Peeters, "A Large set of audio features for sound description (similarity and classification) in the CUIDADO Project," IRCAM, Paris, France, Apr. 2004, Tech. Rep..
[126] S. Essid, "Classification automatique des signaux audio-fréquences: Reconnaissance des instruments de musique," Ph.D. dissertation, Univ. Pierre et Marie Curie, Paris, France, Dec. 2005.
[127] A. Eronen, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs," in *Proc. th Int. Symp. Signal Process. and Its Applicat.*, Paris, France, 2003, pp. 133–136.
[128] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Musical instrument recognizer "instrogram" and its application to music retrieval based on instrument similarity," in *Proc. IEEE Int. Symp. Multimedia*, San Diego, CA, 2006, pp. 265–272.
[129] D. Little and B. Pardo, "Learning musical instruments from mixtures of audio with weak labels," in *Proc. 9th Int. Symp. Music Inf. Retrieval (ISMIR)*, Philadelphia, PA, 2008.
[130] J. Eggink and G. J. Brown, "A missing feature approach to instrument identification in polyphonic music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, China, 2003, pp. 553–556.
[131] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.
[132] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.
[133] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps," *EURASIP J. Appl. Signal Process.*, vol. 2007, pp. 1–15, 2007.
[134] B. Kostek, "Musical instrument recognition and duet analysis employing music information retrieval techniques," *Proc. IEEE*, vol. 92, pp. 712–729, 2004.
[135] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic instrument recognition using spectral clustering," in *Proc. 8th Int. Symp. Music Inf. Retrieval*, Vienna, Austria, 2007.
[136] J. J. Burred, A. Röbel, and T. Sikora, "Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, 2009, pp. 173–176.
[137] , D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York: Wiley/IEEE Press, 2006.
[138] E. Vincent and X. Rodet, "Instrument identification in solo and ensemble music using independent subspace analysis," in *Proc. 5th Int. Symp. Music Inf. Retrieval*, Barcelona, Spain, 2004, pp. 576–581.
[139] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 116–128, Jan. 2008.
[140] K. Kashino and H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Commun.*, vol. 27, pp. 337–349, 1999.
[141] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2006, ch. 6, pp. 267–296.
[142] P. Comon and C. Jutten, *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. New York: Academic, Elsevier, 2010.
[143] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford Univ. , Stanford, CA, 1989.
[144] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 529–540, Mar. 2008.
[145] F. B. Laurent Benaroya and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 191–199, Jan. 2006.
[146] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[147] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[148] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[149] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization,". Whistler, Canada, 2006, extended abstract.

[150] R. G. Emmanuel Vincent and C. Féevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2005.

[151] W. Hess, *Pitch Determination of Speech Signals*. Berlin, Germany: Springer-Verlag, 1983.

[152] W. Hess, *Pitch and Voicing Determination*. New York: Marcel Dekker, 1992, pp. 3–48.

[153] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Amer. (JASA)*, vol. 43, no. 4, pp. 829–834, 1968.

[154] P. Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1982, pp. 180–183.

[155] J.-S. Liénard, F. Signol, and C. Barras, "Speech fundamental frequency estimation using the alternate comb," in *Proc. Interspeech*, Antwerpen, Belgium, 2007.

[156] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer. (JASA)*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[157] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, ch. 14, pp. 495–518.

[158] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[159] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *J. Acoust. Soc. Amer. (JASA)*, vol. 89, no. 6, pp. 2866–2882, 1991.

[160] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, pp. 53–56.

[161] A. de Cheveigne, "Multiple f0 estimation," in *Computational Auditory Scene Analysis*, D. Wang and G. J. Brown, Eds. New York: Wiley/ IEEE Press, 2006.

[162] *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2006.

[163] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," in *Synthesis Lectures on Speech Audio Process.*. San Rafael, CA: Morgan and Claypool, 2009.

[164] R. Badeau, V. Emiya, and B. David, "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra," in *Proc. ICASSP'09*, Taipei, Taiwan, Apr. 2009, pp. 3073–3076.

[165] C. Yeh, "Multiple fundamental frequency estimation of polyphonic recordings," Ph.D. dissertation, Univ. Pierre et Marie Curie (Paris 6), Paris, France, 2008.

[166] F. Signol, C. Barras, and J.-S. Lienard, "Evaluation of the pitch estimation algorithms in the monopitch and multipitch cases," in *Proc. Acoustics'08*, 2008, vol. 123, no. 5, p. 3077.

[167] ***[AUTHOR: Please provide page range]***G. E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, 2007.

[168] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 439–449, Jun. 2004.

[169] J. P. Bello, L. Daudet, and M. Sandler, "Automatic piano transcription using frequency and time-domain information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2242–2251, Aug. 2006.

[170] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2003, pp. 177–180.

[171] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

[172] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[173] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[174] M. Ryynänen and A. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Comput. Music J.*, vol. 32, no. 3, pp. 72–86, 2008.

[175] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 255–266, Feb. 2008.

[176] C. Yeh and A. Roebel, "The expected amplitude of overlapping partials of harmonic sounds," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, Taipei, Taiwan, 2009, pp. 316–319.

[177] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discriminiation and resynthesis," in *Proc. 7th Int. Conf. Digital Audio Effects (DAFX-04)*, Naples, Italy, Oct. 2004.

[178] S. Pauws, "CubyHum: A fully operational query by humming system," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Paris, France, 2002.

[179] J. Weil, J.-L. Durrieu, G. Richard, and T. Sikora, "Automatic generation of lead sheets from polyphonic music signals," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Kobe, Japan, Oct. 2009, pp. 603–608.

[180] R. Foucard, J.-L. Durrieu, M. Lagrange, and G. Richard, "Multimodal similarity between musical streams for cover version detection," in *Proc. ICASSP*, Dallas, TX, Mar. 2010, pp. 5514–5517.

[181] R. P. Paiva, "Melody detection in polyphonic audio," Ph.D. dissertation, Univ. of Coimbra, Coimbra, Portugal, 2006.

[182] M. Goto, "A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2000, vol. 2, pp. 757–760.

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

[184] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," in *Proc. 9th Int. Conf. Digital Audio Effects (DAfX-06)*, 2006, pp. 18–20.

[185] K. Dressler, "Audio melody extraction," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR): Late Breaking Session*, 2010.

[186] J.-L. Durrieu, "Transcription et séparation automatique de la mélodie principale dans les signaux de musique polyphoniques," Ph.D. dissertation, Télécom ParisTech, Paris, France, May 2010.

[187] E. Tsunoo, T. Akase, N. Ono, and S. Sagayama, "Music mood classification by rhythm and bass-line unit pattern analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2010, pp. 265–268.

[188] M. Goto and S. Hayamizu, "A real-time music scene description system: Detecting melody and bass lines in audio signals," in *Proc. Int. Workshop Computat. Auditory Scene Anal.*, Aug. 1999.

[189] S. Hainsworth and M. D. Macleod, "Automatic bass line transcription from polyphonic music," in *Proc. Int. Comput. Music Conf.*, Havana, Cuba, 2001.

[190] M. Ryynänen and A. Klapuri, "Automatic bass line transcription from streaming polyphonic audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Honolulu, HI, 2007, pp. 1437–1440.

[191] A. Kapur, M. Benning, and G. Tzanetakis, "Query-by-beat-boxing: Music retrieval for the dj," in *Proc. 6th Int. Conf. Music Inf. Retrieval, (ISMIR)*, Barcelona, Spain, Sep. 2004.

[192] T. Nakano, J. Ogata, M. Goto, and Y. Hiraga, "A drum pattern retrieval method by voice percussion," in *Proc. 6th Int. Conf. Music Inf. Retrieval, (ISMIR)*, 2004, pp. 550–553.

[193] J. C. C. Chen and A. L. Chen, "Query by rhythm: An approach for song retrieval in music databases," in *Proc. Int. Workshop Res. Iss. Data Eng.*, Feb. 1998.

[194] O. Gillet and G. Richard, "Drum loops retrieval from spoken queries," *J. Intell. Inf. Syst.—Special Iss. Intell. Multimedia Applicat.*, vol. 24, no. 2/3, pp. 159–177, Mar. 2005.

[195] P. Chordia, "Automatic transcription of solo tabla music," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2006.

[196] P. Chordia and A. Rae, "Tabla gyan: A system for realtime tabla recognition and resynthesis," in *Proc. Int. Comput. Music Conf. (ICMC)*, 2008.

[197] O. Gillet and G. Richard, "Automatic e.g., ing of tabla signals," in *Proc. 4th ISMIR Conf. 2003*, Baltimore, MD, Oct. 2003.

[198] D. FitzGerald and J. Paulus, "Unpitched percussion transcription," in *Signal Processing Methods for Music Transcription*. New York: Springer, 2006, pp. 131–162.
[199] V. Sandvold, F. Gouyon, and P. Herrera, "Percussion classification in polyphonic audio recordings using localized sound models," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Oct. 2004, pp. 537–540.
[200] K. Tanghe, S. Degroeve, and B. D. Baets, "An algorithm for detecting and labeling drum events in polyphonic music," in *Proc. 1st MIREX*, London, U.K., Sep. 2005.
[201] O. Gillet and G. Richard, "Drum track transcription of polyphonic music signals using noise subspace projection," in *Proc. 6th Int. Conf. Music Inf. Retrieval, (ISMIR)*, London, U.K., Sep. 2005.
[202] J. Paulus, "Acoustic modelling of drum sounds with hidden markov models for music transcription," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, pp. 241–244.
[203] A. Zils, F. Pachet, O. Delerue, and F. Gouyon, "Automatic extraction of drum tracks from polyphonic music signals," in *Proc. Int. Conf. Web Delivering of Music (WEDELMUSIC)*, Dec. 2002.
[204] K. Yoshii, M. Goto, and H. G. Okuno, "Automatic drum sound description for real-world music using template adaptation and matching methods," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2004.
[205] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *Proc. Int. Symp. Ind. Compon. Anal. Blind Signal Separat. (ICA)*, Apr. 2003.
[206] C. Uhle and C. Dittmar, "Further steps towards drum transcription of polyphonic music," in *Proc. 11th AES Conv.* , May 2004.
[207] D. FitzGerald, E. Coyle, and B. Lawlor, "Sub-band independent subspace analysis for drum transcription," in *Proc. 5th Int. Conf. Digital Audio Effects (DAFX-02)*, 2002.
[208] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using nonnegative matrix factorization and support vector machine," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2005.
[209] J. Paulus and T. Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2005.
[210] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, pp. 613–616.
[211] R. Badeau, B. David, and R. Boyer, "Eds parametric modeling and tracking of audio signals," in *Proc. 5th Int. Conf. Digital Audio Effects (DAFX-02)*, 2002, pp. 26–28.
[212] D. P. W. Ellis, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds. New York: Wiley/IEEE Press, 2006, ch. 4, pp. 115–146.
[213] J. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2006, pp. 314–319.
[214] V. Konz, M. Müller, and S. Ewert, "A multi-perspective evaluation framework for chord recognition," in *Proc. 11th Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 9–14.
[215] M. R. Every and J. E. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1845–1856, Sep. 2006.
[216] R. B. Dannenberg, "Toward automated holistic beat tracking, music analysis and understanding," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005, pp. 366–373.
[217] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1280–1289, Aug. 2010.

Meinard Müller (M'09) received the Diplom degree in mathematics, the Ph.D. degree in computer science, and the Habilitation degree in the field of multimedia retrieval from Bonn University, Bonn, Germany.

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

Daniel P. W. Ellis (M'96–SM'04) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He is an Associate Professor in the Electrical Engineering Department, Columbia University, New York. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with all aspects of extracting high-level information from audio, including speech recognition, music description, and environmental sound processing. He also runs the AUDITORY e-mail list of 1700 worldwide researchers in perception and cognition of sound. He worked at MIT, where he was a Research Assistant at the Media Lab, and he spent several years as a Research Scientist at the International Computer Science Institute, Berkeley, CA, where he remains an external fellow.

Anssi Klapuri (M'06) received the Ph.D. degree from Tampere University of Technology (TUT), Tampere, Finland, in 2004.

He visited as a Post-Doctoral Researcher at the Ecole Centrale de Lille, Lille, France, and Cambridge University, Cambridge, U.K., in 2005 and 2006, respectively. He worked until 2009 as a Professor at TUT. In December 2009, he joined Queen Mary, University of London, London, U.K., as a Lecturer in sound and music processing. His research interests include audio signal processing, auditory modeling, and machine learning.

Gaël Richard (M'02–SM'06) received the State Engineering degree from TELECOM ParisTech (formerly ENST), Paris, France, in 1990, the Ph.D. degree in speech synthesis from LIMSI-CNRS, University of Paris-XI, in 1994 and the *Habilitation à Diriger des Recherches* degree from the University of Paris XI in September 2001.

After the Ph.D. degree, he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches to speech production. Between 1997 and 2001, he successively worked for Matra Nortel Communications, Bois d'Arcy, France, and for Philips Consumer Communications, Montrouge, France. In particular, he was the Project Manager of several large-scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined the Department of Signal and Image Processing, TELECOM ParisTech, where he is now a Full Professor in audio signal processing and Head of the Audio, Acoustics, and Waves Research Group. He is coauthor of over 80 papers, inventor of a number of patents, and one of the experts of the European commission in the field of speech and audio signal processing.

Prof. Richard is a member of EURASIP and an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

Signal Processing for Music Analysis

Meinard Müller, *Member, IEEE*, Daniel P. W. Ellis, *Senior Member, IEEE*, Anssi Klapuri, *Member, IEEE*, and Gaël Richard, *Senior Member, IEEE*

***Abstract*—Music signal processing may appear to be the junior relation of the large and mature field of speech signal processing, not least because many techniques and representations originally developed for speech have been applied to music, often with good results. However, music signals possess specific acoustic and structural characteristics that distinguish them from spoken language or other nonmusical signals. This paper provides an overview of some signal analysis techniques that specifically address musical dimensions such as melody, harmony, rhythm, and timbre. We will examine how particular characteristics of music signals impact and determine these techniques, and we highlight a number of novel music analysis and retrieval tasks that such processing makes possible. Our goal is to demonstrate that, to be successful, music audio signal processing techniques must be informed by a deep and thorough insight into the nature of music itself.**

***Index Terms*—Beat, digital signal processing, harmony, melody, music analysis, music information retrieval, music signals, pitch, rhythm, source separation, timbre, voice separation.**

I. Introduction

MUSIC is a ubiquitous and vital part of the lives of billions of people worldwide. Musical creations and performances are among the most complex and intricate of our cultural artifacts, and the emotional power of music can touch us in surprising and profound ways. Music spans an enormous range of forms and styles, from simple, unaccompanied folk songs, to orchestras and other large ensembles, to a minutely constructed piece of electronic music resulting from months of work in the studio.

The revolution in music distribution and storage brought about by personal digital technology has simultaneously fueled tremendous interest in and attention to the ways that information technology can be applied to this kind of content. From browsing personal collections, to discovering new artists, to managing and protecting the rights of music creators, computers are now deeply involved in almost every aspect of music consumption, which is not even to mention their vital role in much of today's music production.

This paper concerns the application of signal processing techniques to music signals, in particular to the problems of analyzing an existing music signal (such as piece in a collection) to extract a wide variety of information and descriptions that may be important for different kinds of applications. We argue that there is a distinct body of techniques and representations that are molded by the particular properties of music audio—such as the pre-eminence of distinct fundamental periodicities (pitches), the preponderance of overlapping sound sources in musical ensembles (polyphony), the variety of source characteristics (timbres), and the regular hierarchy of temporal structures (beats). These tools are more or less unlike those encountered in other areas of signal processing, even closely related fields such as speech signal processing. In any application, the more closely the processing can reflect and exploit the particular properties of the signals at hand, the more successful it will be. Musical signals, despite their enormous diversity, do exhibit a number of key properties that give rise to the techniques of music signal processing, as we shall see.

The application of signal processing to music signals is hardly new, of course. It could be argued to be the basis of the theremin, a 1920s instrument in which an oscillator is controlled by the capacitance of the player's hands near its antennae. The development of modern signal processing in the 1940s and 1950s led directly the first wave of electronic music, in which composers such as Karlheinz Stockhausen created music using signal generators, ring modulators, etc., taken straight from electronics labs. Following the advent of general-purpose digital computers in the 1960s and 1970s, it was not long before they were used to synthesize music by pioneers like Max Matthews and John Pierce. Experimental music has remained a steady source of innovative applications of signal processing, and has spawned a significant body of sophisticated techniques for synthesizing and modifying sounds [1], [2].

As opposed to synthesis, which takes a compact, abstract description such as a musical score and creates a corresponding signal, our focus is analysis—for instance, the inverse problem of recovering a score-level description given only the audio. It turns out that this problem is very computationally demanding, and although efforts at automatic transcription, for example, date back to the mid 1970s [3], the vastly improved computational resources of recent years, along with the demands and opportunities presented by massive online music collections, have led to a recent explosion in this research. The first International Symposium on Music Information Retrieval[1] was held in 2000; this annual meeting is now a thriving interdisciplinary community with over 100 papers presented at the 2010 conference in Utrecht.

Manuscript received September 27, 2010; accepted December 03, 2010. Date of publication February 04, 2011; date of current version nulldate. The work of M. Müller was supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI). The associate editor coordinating the review of this manuscript and approving it for publication was XXXX XXXXXX.

M. Müller is with the Saarland University, 66123 Saarbrücken, Germany, and also with the Max-Planck Institut für Informatik, 66123 Saarbrücken, Germany (e-mail: meinard@mpi-inf.mpg.de).

D. P. W. Ellis is with Electrical Engineering Department, Columbia University, New York, NY 10027 USA (e-mail: dpwe@ee.columbia.edu).

A. Klapuri is with Queen Mary University of London, London E1 4NS, U.K. (e-mail: anssi.klapuri@elec.qmul.ac.uk).

G. Richard is with the Institut TELECOM, Télécom ParisTech, CNRS-LTCI, F-75634 Paris Cedex 13, France (e-mail: gael.richard@telecom-paristech.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2011.2112333

[1][Online]. Available: http://www.ismir.net/.

In the rest of the paper, we will present several of the distinctive aspects of the musical signal, and the most significant approaches that have been developed for its analysis. While many techniques are initially borrowed from speech processing or other areas of signal processing, the unique properties and stringent demands of music signals have dictated that simple repurposing is not enough, leading to some inspired and elegant solutions. We have organized the material along particular musical dimensions: In Section II, we discuss the nature of pitch and harmony in music, and present time–frequency representations used in their analysis. Then, in Section III, we address the musical aspects of note onsets, beat, tempo, and rhythm. In Section IV, we discuss models representing the timbre and instrumentation of music signals and introduce various methods for recognizing musical instruments on audio recordings. Finally, in Section V, we show how acoustic and musical characteristics can be utilized to separate musical voices, such as the melody and bass line, from polyphonic music. We conclude in Section VI with a discussion of open problems and future directions.

II. Pitch and Harmony

Pitch is a ubiquitous feature of music. Although a strict definition of music is problematic, the existence of sequences of sounds with well-defined fundamental periods—i.e., individual notes with distinct pitches—is a very common feature. In this section, we discuss the objective properties of musical pitch and its use to create musical harmony, then go on to present some of the more common time–frequency representations used in music signal analysis. A number of music applications based on these representations are described.

A. Musical Pitch

Most musical instruments—including string-based instruments such as guitars, violins, and pianos, as well as instruments based on vibrating air columns such as flutes, clarinets, and trumpets—are explicitly constructed to allow performers to produce sounds with easily controlled, locally stable fundamental periods. Such a signal is well described as a harmonic series of sinusoids at multiples of a fundamental frequency, and results in the percept of a musical note (a single perceived event) at a clearly defined pitch in the mind of the listener. With the exception of unpitched instruments like drums, and a few inharmonic instruments such as bells, the periodicity of individual musical notes is rarely ambiguous, and thus equating the perceived pitch with fundamental frequency is common.

Music exists for the pleasure of human listeners, and thus its features reflect specific aspects of human auditory perception. In particular, humans perceive two signals whose fundamental frequencies fall in a ratio 2:1 (an octave) as highly similar [4] (sometimes known as "octave equivalence"). A sequence of notes—a melody—performed at pitches exactly one octave displaced from an original will be perceived as largely musically equivalent. We note that the sinusoidal harmonics of a fundamental at f_0 at frequencies $f_0, 2f_0, 3f_0, 4f_0, \ldots$ are a proper superset of the harmonics of a note with fundamental $2f_0$ (i.e., $2f_0, 4f_0, 6f_0, \ldots$), and this is presumably the basis of the perceived similarity. Other pairs of notes with frequencies in simple ratios, such as f_0 and $3f_0/2$ will also share many harmonics, and are also perceived as similar—although not as close as the octave. Fig. 1 shows the waveforms and spectrograms of middle C (with fundamental frequency 262 Hz) played on a piano and a violin. Zoomed-in views above the waveforms show the relatively stationary waveform with a 3.8-ms period in both cases. The spectrograms (calculated with a 46-ms window) show the harmonic series at integer multiples of the fundamental. Obvious differences between piano and violin sound include the decaying energy within the piano note, and the slight frequency modulation ("vibrato") on the violin.

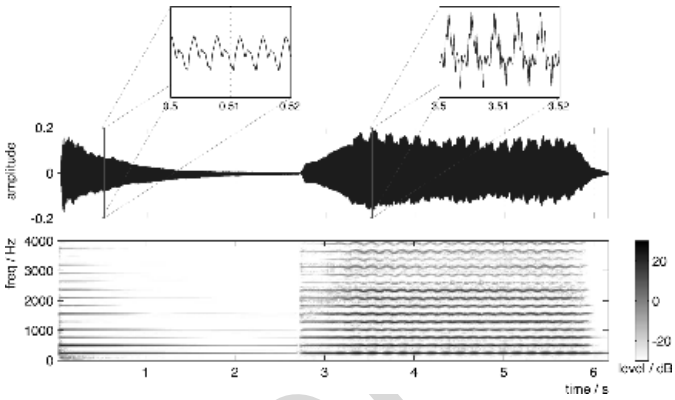

Fig. 1. Middle C (262 Hz) played on a piano and a violin. The top pane shows the waveform, with the spectrogram below. Zoomed-in regions shown above the waveform reveal the 3.8-ms fundamental period of both notes.

Although different cultures have developed different musical conventions, a common feature is the musical "scale," a set of discrete pitches that repeats every octave, from which melodies are constructed. For example, contemporary western music is based on the "equal tempered" scale, which, by a happy mathematical coincidence, allows the octave to be divided into twelve equal steps on a logarithmic axis while still (almost) preserving intervals corresponding to the most pleasant note combinations. The equal division makes each frequency $2^{1/12} \approx 1.06\times$ larger than its predecessor, an interval known as a semitone. The coincidence is that it is even possible to divide the octave uniformly into such a small number of steps, and still have these steps give close, if not exact, matches to the simple integer ratios that result in consonant harmonies, e.g., $\left(2^{1/12}\right)^7 = 1.498 \approx 3/2$, and $\left(2^{1/12}\right)^5 = 1.335 \approx 4/3$. The western major scale spans the octave using seven of the twelve steps—the "white notes" on a piano, denoted by C, D, E, F, G, A, B. The spacing between successive notes is two semitones, except for E/F and B/C which are only one semitone apart. The "black notes" in between are named in reference to the note immediately below (e.g., C♯), or above (D♭), depending on musicological conventions. The octave degree denoted by these symbols is sometimes known as the pitch's *chroma*, and a particular pitch can be specified by the concatenation of a chroma and an octave number (where each numbered octave spans C to B). The lowest note on a piano is A0 (27.5 Hz), the highest note is C8 (4186 Hz), and middle C (262 Hz) is C4.

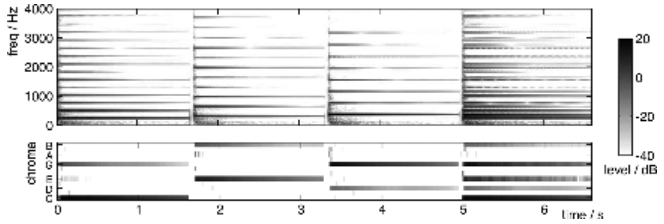

Fig. 2. Middle C, followed by the E and G above, then all three notes together—a C Major triad—played on a piano. Top pane shows the spectrogram; bottom pane shows the chroma representation.

B. Harmony

While sequences of pitches create melodies—the "tune" of a music, and the only part reproducible by a monophonic instrument such as the voice—another essential aspect of much music is harmony, the simultaneous presentation of notes at different pitches. Different combinations of notes result in different musical colors or "chords," which remain recognizable regardless of the instrument used to play them. Consonant harmonies (those that sound "pleasant") tend to involve pitches with simple frequency ratios, indicating many shared harmonics. Fig. 2 shows middle C (262 Hz), E (330 Hz), and G (392 Hz) played on a piano; these three notes together form a C Major triad, a common harmonic unit in western music. The figure shows both the spectrogram and the chroma representation, described in Section II-E below. The ubiquity of simultaneous pitches, with coincident or near-coincident harmonics, is a major challenge in the automatic analysis of music audio: note that the chord in Fig. 2 is an unusually easy case to visualize thanks to its simplicity and long duration, and the absence of vibrato in piano notes.

C. Time–Frequency Representations

Some music audio applications, such as transcribing performances, call for explicit detection of the fundamental frequencies present in the signal, commonly known as pitch tracking. Unfortunately, the presence of multiple, simultaneous notes in polyphonic music renders accurate pitch tracking very difficult, as discussed further in Section V. However, there are many other applications, including chord recognition and music matching, that do not require explicit detection of pitches, and for these tasks several representations of the pitch and harmonic information—the "tonal content" of the audio—commonly appear. Here, we introduce and define these basic descriptions.

As in other audio-related applications, the most popular tool for describing the time-varying energy across different frequency bands is the short-time Fourier transform (STFT), which, when visualized as its magnitude, is known as the spectrogram (as in Figs. 1 and 2). Formally, let x be a discrete-time signal obtained by uniform sampling of a waveform at a sampling rate of F_s Hz. Using an N-point tapered window w (e.g., Hamming $w(n) = 0.54 - 0.46\cos(2\pi n/N)$ for $n \in [0:N-1] := \{0,1,\ldots N-1\}$) and an overlap of half a window length, we obtain the STFT

$$X(t,k) = \sum_{n=0}^{N-1} w(n)x(n+t\cdot N/2)\exp\{-j2\pi kn/N\} \quad (1)$$

with $t \in [0:T-1]$ and $k \in [0:K]$. Here, T determines the number of frames, $K = N/2$ is the index of the last unique frequency value, and thus $X(t,k)$ corresponds to the window beginning at time $t \cdot N/(2F_s)$ in seconds and frequency

$$f_{\text{coeff}}(k) = \frac{k}{N}\cdot F_s \quad (2)$$

in Hertz (Hz). Typical values of $F_s = 44\,100$ and $N = 4096$ give a window length of 92.8 ms, a time resolution of 46.4 ms, and frequency resolution of 10.8 Hz.

$X(t,k)$ is complex-valued, with the phase depending on the precise alignment of each short-time analysis window. Often it is only the magnitude $|X(t,k)|$ that is used. In Figs. 1 and 2, we see that this spectrogram representation carries a great deal of information about the tonal content of music audio, even in Fig. 2's case of multiple, overlapping notes. However, close inspection of the individual harmonics at the right-hand side of that figure—at 780 and 1300 Hz, for example—reveals amplitude modulations resulting from phase interactions of close harmonics, something that cannot be exactly modeled in a magnitude-only representation.

D. Log-Frequency Spectrogram

As mentioned above, our perception of music defines a logarithmic frequency scale, with each doubling in frequency (an octave) corresponding to an equal musical interval. This motivates the use of time–frequency representations with a similar logarithmic frequency axis, which in fact correspond more closely to representation in the ear [5]. (Because the bandwidth of each bin varies in proportion to its center frequency, these representations are also known as "constant-Q transforms," since each filter's effective center frequency-to-bandwidth ratio—its Q—is the same.) With, for instance, 12 frequency bins per octave, the result is a representation with one bin per semitone of the equal-tempered scale.

A simple way to achieve this is as a mapping applied to an STFT representation. Each bin in the log-frequency spectrogram is formed as a linear weighting of corresponding frequency bins from the original spectrogram. For a log-frequency axis with K_L bins, this calculation can be expressed in matrix notation as $\mathbf{Y} = \mathbf{M}\mathbf{X}$, where $\mathbf{Y}$ is the log-frequency spectrogram with K_L rows and T columns, $\mathbf{X}$ is the original STFT magnitude array $|X(t,k)|$ (with t indexing columns and k indexing rows). $\mathbf{M}$ is a weighting matrix consisting of K_L rows, each of $K+1$ columns, that give the weight of STFT bin $|X(\cdot,k)|$ contributing to log-frequency bin $Y(\cdot,\ell)$. For instance, using a Gaussian window

$$M(\ell,k) = \exp\left\{-\frac{1}{2B^2}\left(\log_2\frac{f_{\text{coeff}}(k)}{f_{\min}} - \frac{\ell}{N_O}\right)^2\right\} \quad (3)$$

where B defines the bandwidth of the filterbank as the frequency difference (in octaves) at which the bin has fallen to $\exp(-1/2)$ of its peak gain. $f_{\min}$ is the frequency of the lowest bin ($\ell = 0$) and N_O is the number of bins per octave in the log-frequency axis. The calculation is illustrated in Fig. 3, where the top-left image is the matrix $\mathbf{M}$, the top right is the conventional spectrogram $\mathbf{X}$, and the bottom right shows the resulting log-frequency spectrogram $\mathbf{Y}$.

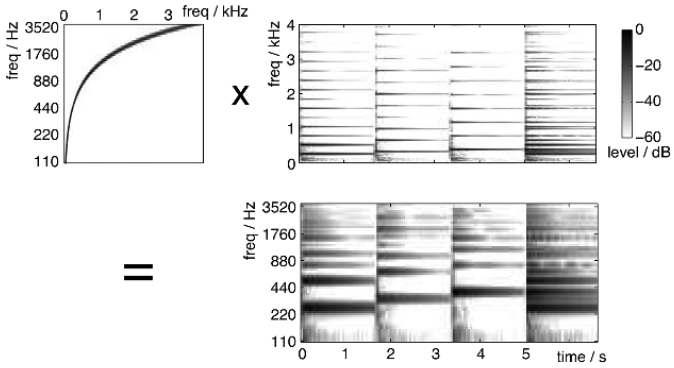

Fig. 3. Calculation of a log-frequency spectrogram as a columnwise linear mapping of bins from a conventional (linear-frequency) spectrogram. The images may be interpreted as matrices, but note the bottom-to-top sense of the frequency (row) axis.

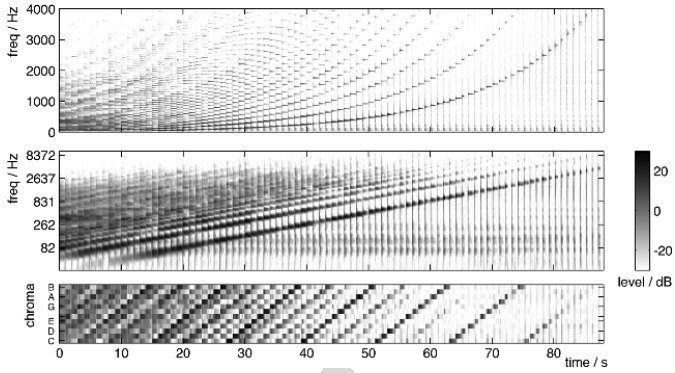

Fig. 4. Three representations of a chromatic scale comprising every note on the piano from lowest to highest. Top pane: conventional spectrogram (93-ms window). Middle pane: log-frequency spectrogram (186-ms window). Bottom pane: chromagram (based on 186-ms window).

Although conceptually simple, such a mapping often gives unsatisfactory results: in the figure, the logarithmic frequency axis uses $N_O = 12$ (one bin per semitone), starting at $f_{\min} = 110$ Hz (A2). At this point, the log-frequency bins have centers only 6.5 Hz apart; to have these centered on distinct STFT bins would require a window of 153 ms, or almost 7000 points at $F_s = 44\,100$ Hz. Using a 64-ms window, as in the figure, causes blurring of the low-frequency bins. Yet, by the same token, the highest bins shown—five octaves above the lowest—involve averaging together many STFT bins.

The long time window required to achieve semitone resolution at low frequencies has serious implications for the temporal resolution of any analysis. Since human perception of rhythm can often discriminate changes of 10 ms or less [4], an analysis window of 100 ms or more can lose important temporal structure. One popular alternative to a single STFT analysis is to construct a bank of individual bandpass filters, for instance one per semitone, each tuned the appropriate bandwidth and with minimal temporal support [6, Sec. 3.1]. Although this loses the famed computational efficiency of the fast Fourier transform, some of this may be regained by processing the highest octave with an STFT-based method, downsampling by a factor of 2, then repeating for as many octaves as are desired [7], [8]. This results in different sampling rates for each octave of the analysis, raising further computational issues. A toolkit for such analysis has been created by Schorkhuber and Klapuri [9].

E. Time-Chroma Representations

Some applications are primarily concerned with the chroma of the notes present, but less with the octave. Foremost among these is chord transcription—the annotation of the current chord as it changes through a song. Chords are a joint property of all the notes sounding at or near a particular point in time, for instance the C Major chord of Fig. 2, which is the unambiguous label of the three notes C, E, and G. Chords are generally defined by three or four notes, but the precise octave in which those notes occur is of secondary importance. Thus, for chord recognition, a representation that describes the chroma present but "folds" the octaves together seems ideal. This is the intention of chroma representations, first introduced as Pitch Class Profiles in [10]; the description as chroma was introduced in [11].

A typical chroma representation consists of a 12-bin vector for each time step, one for each chroma class from C to B. Given a log-frequency spectrogram representation with semitone resolution from the preceding section, one way to create chroma vectors is simply to add together all the bins corresponding to each distinct chroma [6, Ch. 3]. More involved approaches may include efforts to include energy only from strong sinusoidal components in the audio, and exclude non-tonal energy such as percussion and other noise. Estimating the precise frequency of tones in the lower frequency range may be important if the frequency binning of underlying transform is not precisely aligned to the musical scale [12].

Fig. 4 shows a chromatic scale, consisting of all 88 piano keys played one a second in an ascending sequence. The top pane shows the conventional, linear-frequency spectrogram, and the middle pane shows a log-frequency spectrogram calculated as in Fig. 3. Notice how the constant ratio between the fundamental frequencies of successive notes appears as an exponential growth on a linear axis, but becomes a straight line on a logarithmic axis. The bottom pane shows a 12-bin chroma representation (a "chromagram") of the same data. Even though there is only one note sounding at each time, notice that very few notes result in a chroma vector with energy in only a single bin. This is because although the fundamental may be mapped neatly into the appropriate chroma bin, as will the harmonics at $2f_0, 4f_0, 8f_0$, etc. (all related to the fundamental by octaves), the other harmonics will map onto other chroma bins. The harmonic at $3f_0$, for instance, corresponds to an octave plus 7 semitones ($2^{(12+7)/12} \approx 3$), thus for the C4 sounding at 40 s, we see the second most intense chroma bin after C is the G seven steps higher. Other harmonics fall in other bins, giving the more complex pattern. Many musical notes have the highest energy in the fundamental harmonic, and even with a weak fundamental, the root chroma is the bin into which the greatest number of low-order harmonics fall, but for a note with energy across a large number of harmonics—such as the lowest notes in the figure—the chroma vector can become quite cluttered.

One might think that attempting to attenuate higher harmonics would give better chroma representations by reducing these alias terms. In fact, many applications are improved by whitening the spectrum—i.e., boosting weaker bands to make the energy approximately constant across the spectrum. This

helps remove differences arising from the different spectral balance of different musical instruments, and hence better represents the tonal, and not the timbral or instrument-dependent, content of the audio. In [13], this is achieved by explicit normalization within a sliding local window, whereas [14] discards low-order cepstral coefficients as a form of "liftering."

Chroma representations may use more than 12 bins per octave to reflect finer pitch variations, but still retain the property of combining energy from frequencies separated by an octave [15], [16]. To obtain robustness against global mistunings (resulting from instruments tuned to a standard other than the 440 Hz A4, or distorted through equipment such as tape recorders running at the wrong speed), practical chroma analyses need to employ some kind of adaptive tuning, for instance by building a histogram of the differences between the frequencies of all strong harmonics and the nearest quantized semitone frequency, then shifting the semitone grid to match the peak of this histogram [12]. It is, however, useful to limit the range of frequencies over which chroma is calculated. Human pitch perception is most strongly influenced by harmonics that occur in a "dominance region" between about 400 and 2000 Hz [4]. Thus, after whitening, the harmonics can be shaped by a smooth, tapered frequency window to favor this range.

Notwithstanding the claims above that octave is less important than chroma, the lowest pitch in a collection of notes has a particularly important role in shaping the perception of simultaneous notes. This is why many musical ensembles feature a "bass" instrument—the double-bass in an orchestra, or the bass guitar in rock music—responsible for playing very low notes. As discussed in Section V, some applications explicitly track a bass line, but this hard decision can be avoided be avoided by calculating a second chroma vector over a lower frequency window, for instance covering 50 Hz to 400 Hz [17].

Code toolboxes to calculate chroma features are provided by Ellis[2] and Müller.[3]

F. Example Applications

Tonal representations—especially chroma features—have been used for a wide range of music analysis and retrieval tasks in which it is more important to capture polyphonic musical content without necessarily being concerned about the instrumentation. Such applications include chord recognition, alignment, "cover song" detection, and structure analysis.

1) Chord Recognition: As discussed above, chroma features were introduced as Pitch Class Profiles in [10], specifically for the task of recognizing chords in music audio. As a global property of the current notes and context, chords can be recognized based on a global representation of a short window. Moreover, shifting notes up or down by an octave rarely has much impact on the chord identity, so the octave-invariant properties of the chroma vector make it particularly appropriate. There has been a large amount of subsequent work on chord recognition, all based on some variant of chroma [15], [18]–[22]. Developments have mainly focused on the learning and classification aspects of the system: for instance, [15] noted the direct analogy between music audio with transcribed chord sequences (e.g., in "real books") that lack exact temporal alignments, and the speech recordings with unaligned word transcriptions used to train speech recognitions: they used the same Baum–Welch procedure to simultaneously estimate both models for the features of each chord, and the label alignment in the training data. Although later work has been able to take advantage of an increasing volume of manually labeled chord transcriptions [23], significant benefits have been attributed to refinements in feature extraction including separation of tonal (sinusoidal) and percussive (transient or noisy) energy [22], and local spectral normalization prior to chroma calculation [13].

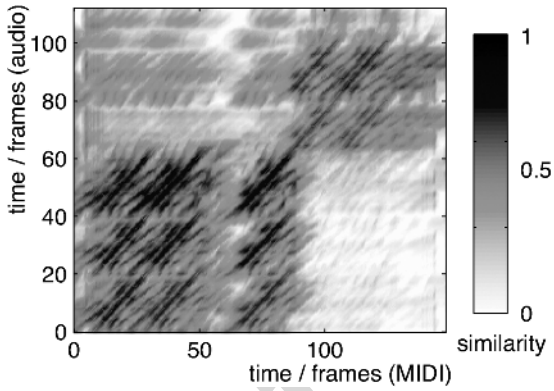

Fig. 5. Similarity matrix comparing a MIDI version of "And I Love Her" (horizontal axis) with the original Beatles recording (vertical axis). From [24].

2) Synchronization and Alignment: A difficult task such as chord recognition or polyphonic note transcription can be made substantially easier by employing an existing, symbolic description such as a known chord sequence, or the entire musical score. Then, the problem becomes that of *aligning* or *synchronizing* the symbolic description to the music audio [6, Ch. 5], making possible innovative applications such as an animated musical score display that synchronously highlights the appropriate bar in time with the music [25]. The core of such an application is to align compatible representations of each component with an efficient technique such as Dynamic Time Warping (DTW) [26], [27]. A time-chroma representation can be directly predicted from the symbolic description, or an electronic score such as a MIDI file can be synthesized into audio, then the synthesized form itself analyzed for alignment to the original recording [28], [29]. Fig. 5 shows a similarity matrix comparing a MIDI version of "And I Love Her" by the Beatles with the actual recording [24]. A similarity matrix $\mathbf{M_{sim}}(t_a, t_b)$ is populated by some measure of similarity $S(\mathbf{X_a}(t_a), \mathbf{X_b}(t_b))$ between representation $\mathbf{X_a}$ at time t_a and $\mathbf{X_b}$ at t_b. For instance, if both $\mathbf{X_a}$ and $\mathbf{X_b}$ are chroma representations, S could be the normalized correlation, $S(\mathbf{X}, \mathbf{Y}) = \mathbf{X} \cdot \mathbf{Y}/(|\mathbf{X}||\mathbf{Y}|)$. Differences in structure and timing between the two versions are revealed by wiggles in the dark ridge closest to the leading diagonal; flanking ridges relate to musical structure, discussed below.

A similar approach can synchronize different recordings of the same piece of music, for instance to allow switching, in real-time, between performances of a piano sonata by different

[2][Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/chroma-ansyn/.

[3][Online]. Available: http://www.mpi-inf.mpg.de/~mmueller/chromatoolbox/.

pianists [25], [30], [31]. In this case, the relatively poor temporal accuracy of tonal representations may be enhanced by the addition of features better able to achieve precise synchronization of onset events [32].

3) "Cover Song" Detection: In popular music, an artist may record his or her own version of another artist's composition, often incorporating substantial changes to instrumentation, tempo, structure, and other stylistic aspects. These alternate interpretations are sometimes known as "cover" versions, and present a greater challenge to alignment, due to the substantial changes. The techniques of chroma feature representation and DTW are, however, still the dominant approaches over several years of development and formal evaluation of this task within the MIREX campaign [33]. Gross structural changes will interfere with conventional DTW, so it must either be modified to report "local matches" [34], or replaced by a different technique such as cross-correlation of beat-synchronous chroma representations [12]. The need for efficient search within very large music collections can be satisfied with efficient hash-table representation of the music broken up into smaller fragments [35]–[37].

4) Structure Recovery: A similarity matrix that compares a piece of music to itself will have a perfectly straight leading diagonal ridge, but will likely have flanking ridges similar to those visible in Fig. 5. These ridges indicate that a certain portion of the signal resembles an earlier (or later) part—i.e., the signal exhibits some repetitive structure. Recurring melodies and chord sequences are ubiquitous in music, which frequently exhibits a hierarchical structure. In popular music, for instance, the song may consist of an introduction, a sequence of alternating verse and chorus, a solo or bridge section, etc. Each of these segments may in turn consist of multiple phrases or lines with related or repeating structure, and the individual phrases may themselves consist of repeating or nearly repeating patterns of notes. [38], for instance, argues that the observation and acquisition of this kind of structure is an important part of the enjoyment of music listening.

Automatic segmentation and decomposition according to this structure is receiving an increasing level of attention; see [39] for a recent review. Typically, systems operate by 1) finding off-diagonal ridges in a similarity matrix to identify and segment into repeating phrases [40], [41], and/or 2) finding segmentation points such that some measure of statistical similarity is maximized within segments, but minimized between adjacent segments [42], [43]. Since human labelers exhibit less consistency on this annotation tasks than for, say, beat or chord labeling, structure recovery is sometimes simplified into problems such as identifying the "chorus," a frequently repeated and usually obvious part of popular songs [11]. Other related problems include searching for structures and motifs that recur within and across different songs within a given body of music [44], [45].

III. Tempo, Beat, and Rhythm

The musical aspects of tempo, beat, and rhythm play a fundamental role for the understanding of and the interaction with music [46]. It is the *beat*, the steady pulse that drives music forward and provides the temporal framework of a piece of music

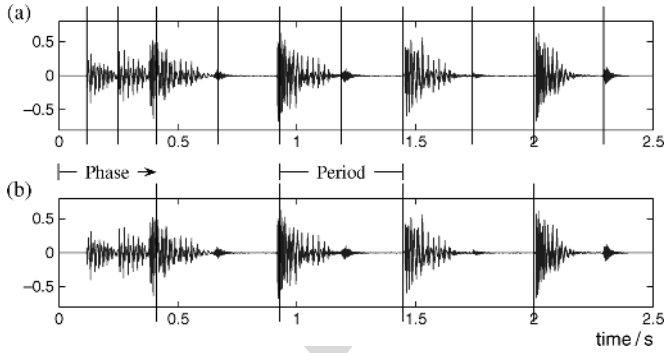

Fig. 6. Waveform representation of the beginning of *Another one bites the dust* by Queen. (a) Note onsets. (b) Beat positions.

[47]. Intuitively, the beat can be described as a sequence of perceived pulses that are regularly spaced in time and correspond to the pulse a human taps along when listening to the music [48]. The term *tempo* then refers to the rate of the pulse. Musical pulses typically go along with note onsets or percussive events. Locating such events within a given signal constitutes a fundamental task, which is often referred to as *onset detection*. In this section, we give an overview of recent approaches for extracting onset, tempo, and beat information from music signals, and then indicate how this information can be applied to derive higher-level rhythmic patterns.

A. Onset Detection and Novelty Curve

The objective of *onset detection* is to determine the physical starting times of notes or other musical events as they occur in a music recording. The general idea is to capture sudden changes in the music signal, which are typically caused by the onset of novel events. As a result, one obtains a so-called *novelty curve*, the peaks of which indicate onset candidates. Many different methods for computing novelty curves have been proposed; see [49] and [50] for an overview. For example, playing a note on a percussive instrument typically results in a sudden increase of the signal's energy, see Fig. 6(a). Having such a pronounced attack phase, note onset candidates may be determined by locating time positions, where the signal's amplitude envelope starts to increase [49]. Much more challenging, however, is the detection of onsets in the case of non-percussive music, where one often has to deal with soft onsets or blurred note transitions. This is often the case for vocal music or classical music dominated by string instruments. Furthermore, in complex polyphonic mixtures, simultaneously occurring events may result in masking effects, which makes it hard to detect individual onsets. As a consequence, more refined methods have to be used for computing the novelty curves, e.g., by analyzing the signal's spectral content [49], [51], pitch [51], [52], harmony [53], [54], or phase [49], [55]. To handle the variety of different signal types, a combination of novelty curves particularly designed for certain classes of instruments can improve the detection accuracy [51], [56]. Furthermore, to resolve masking effects, detection functions were proposed that analyze the signal in a bandwise fashion to extract transients occurring in certain frequency regions of the signal [57], [58]. For example, as a side-effect of

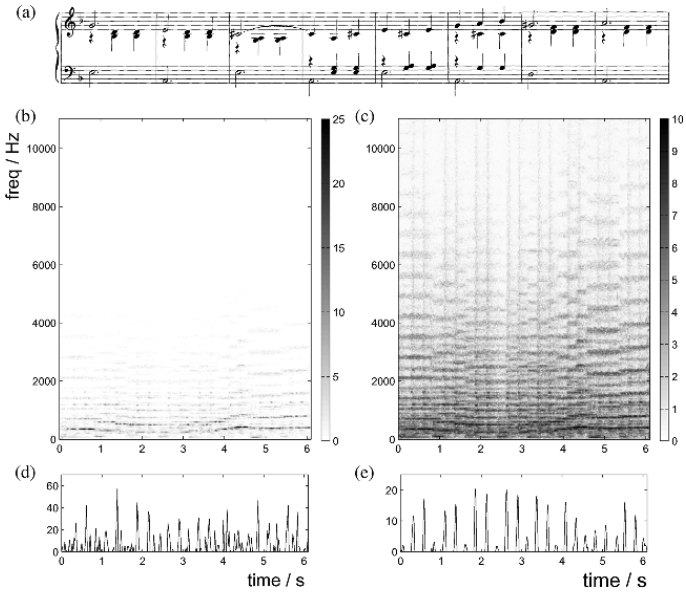

Fig. 7. Excerpt of Shostakovich's Waltz No. 2 from the *Suite for Variety Orchestra No.* 1. (a) Score representation (in a piano reduced version). (b) Magnitude spectrogram. (c) Compressed spectrogram using $C = 1000$. (d) Novelty curve derived from (b). (e) Novelty curve derived from (c).

a sudden energy increase, one can often observe an accompanying broadband noise burst in the signal's spectrum. This effect is mostly masked by the signal's energy in lower frequency regions, but it is well detectable in the higher frequency regions of the spectrum [59]. A widely used approach to onset detection in the frequency domain is the *spectral flux* [49], [60], where changes of pitch and timbre are detected by analyzing the signal's short-time spectrum.

To illustrate some of these ideas, we now describe a typical spectral-based approach for computing novelty curves. Given a music recording, a short-time Fourier transform is used to obtain a spectrogram $X = (X(t,k))_{t,k}$ with $k \in [0 : K]$ and $t \in [0 : T-1]$ as in (1). Note that the Fourier coefficients of X are linearly spaced on the frequency axis. Using suitable binning strategies, various approaches switch over to a logarithmically spaced frequency axis, e.g., by using mel-frequency bands or pitch bands; see [57], [58], and Section II-D. Keeping the linear frequency axis puts greater emphasis on the high-frequency regions of the signal, thus accentuating the aforementioned noise bursts visible as high-frequency content. One simple, yet important step, which is often applied in the processing of music signals, is referred to as *logarithmic compression*; see [57]. In our context, this step consists in applying a logarithm to the magnitude spectrogram $|X|$ of the signal yielding $Y = \log(1+C\cdot|X|)$ for a suitable constant $C > 1$. Such a compression step not only accounts for the logarithmic sensation of human sound intensity, but also balances out the dynamic range of the signal. In particular, by increasing C, low-intensity values in the high-frequency spectrum become more prominent. This effect is clearly visible in Fig. 7, which shows the magnitude spectrogram $|X|$ and the compressed spectrogram Y for a recording of a Waltz by Shostakovich. On the downside, a large compression factor C may also amplify non-relevant low-energy noise components.

To obtain a novelty curve, one basically computes the discrete derivative of the compressed spectrum Y. More precisely, one sums up only positive intensity changes to emphasize onsets while discarding offsets to obtain the novelty function $\Delta : [0 : T-2] \to \mathbb{R}$:

$$\Delta(t) = \sum_{k=0}^{K} |Y(t+1,k) - Y(t,k)|_{\geq 0} \tag{4}$$

for $t \in [0 : T-2]$, where $|x|_{\geq 0} = x$ for a non-negative real number x and $|x|_{\geq 0} = 0$ for a negative real number x. In many implementations, higher order smoothed differentiators are used [61] and the resulting curve is further normalized [62], [63]. Fig. 7(e) shows a typical novelty curve for our Shostakovich example. As mentioned above, one often process the spectrum in a bandwise fashion obtaining a novelty curve for each band separately [57], [58]. These novelty curves are then weighted and summed up to yield a final novelty function.

The peaks of the novelty curve typically indicate the positions of note onsets. Therefore, to explicitly determine the positions of note onsets, one employs peak picking strategies based on fixed or adaptive thresholding [49], [51]. In the case of noisy novelty curves with many spurious peaks, however, this is a fragile and error-prone step. Here, the selection of the relevant peaks that correspond to true note onsets becomes a difficult or even infeasible problem. For example, in the Shostakovich Waltz, the first beats (downbeats) of the 3/4 meter are played softly by non-percussive instruments leading to relatively weak and blurred onsets, whereas the second and third beats are played staccato supported by percussive instruments. As a result, the peaks of the novelty curve corresponding to downbeats are hardly visible or even missing, whereas peaks corresponding to the percussive beats are much more pronounced, see Fig. 7(e).

B. Periodicity Analysis and Tempo Estimation

Avoiding the explicit determination of note onset, novelty curves are often directly analyzed in order to detect reoccurring or quasi-periodic patterns, see [64] for an overview of various approaches. Here, generally speaking, one can distinguish between three different methods. The autocorrelation method allows for detecting periodic self-similarities by comparing a novelty curve with time-shifted (localized) copies [65]–[68]. Another widely used method is based on a bank of comb filter resonators, where a novelty curve is compared with templates that consists of equally spaced spikes covering a range of periods and phases [57], [58]. Third, the short-time Fourier transform can be used to derive a time–frequency representation of the novelty curve [62], [63], [67]. Here, the novelty curve is compared with templates consisting of sinusoidal kernels each representing a specific frequency. Each of the methods reveals periodicity properties of the underlying novelty curve from which one can estimate the tempo or beat structure. The intensities of the estimated periodicity, tempo, or beat properties typically change over time and are often visualized by means of spectrogram-like representations referred to as *tempogram* [69], *rhythmogram* [70], or *beat spectrogram* [71].

Exemplarily, we introduce the concept of a tempogram while discussing two different periodicity estimation methods. Let $[0 : T-1]$ (as for the novelty curve) denote the sampled time axis,

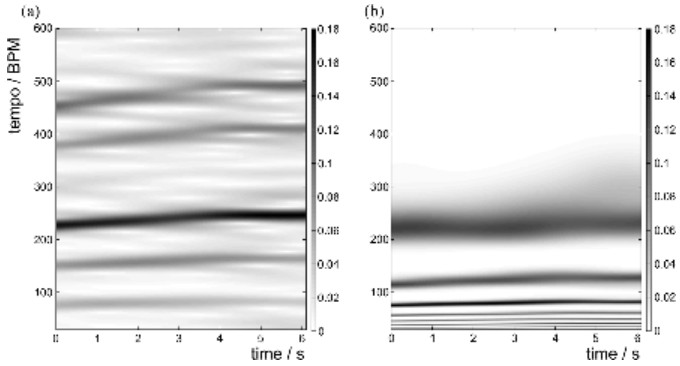

Fig. 8. Excerpt of Shostakovich's Waltz No. 2 from the *Suite for Variety Orchestra No.* 1. (a) Fourier tempogram. (b) Autocorrelation tempogram.

which we extend to $\mathbb{Z}$ to avoid boundary problems. Furthermore, let $\Theta \subset \mathbb{R}_{>0}$ be a set of tempi specified in beats per minute (BPM). Then, a tempogram is mapping $\mathcal{T} : \mathbb{Z} \times \Theta \to \mathbb{R}_{\geq 0}$ yielding a time-tempo representation for a given time-dependent signal. For example, suppose that a music signal has a dominant tempo of $\tau = 220$ BPM around position t, then the corresponding value $\mathcal{T}(t,\tau)$ is large, see Fig. 8. In practice, one often has to deal with tempo ambiguities, where a tempo τ is confused with integer multiples $\tau, 2\tau, 3\tau, \ldots$ (referred to as *harmonics* of τ) and integer fractions $\tau, \tau/2, \tau/3, \ldots$ (referred to as *subharmonics* of τ). To avoid such ambiguities, a mid-level tempo representation referred to as *cyclic tempograms* can be constructed, where tempi differing by a power of two are identified [72], [73]. This concept is similar to the cyclic chroma features, where pitches differing by octaves are identified, cf. Section II-E. We discuss the problem of tempo ambiguity and pulse level confusion in more detail in Section III-C.

A tempogram can be obtained by analyzing a novelty curve Δ with respect to local periodic patterns using a short-time Fourier transform [62], [63], [67]. To this end, one fixes a window function $W : \mathbb{Z} \to \mathbb{R}$ of finite length centered at $t = 0$ (e.g., a centered Hann window of size $2N+1$ for some $N \in \mathbb{N}$). Then, for a frequency parameter $\omega \in \mathbb{R}_{\geq 0}$, the complex Fourier coefficient $\mathcal{F}(t,\omega)$ is defined by

$$\mathcal{F}(t,\omega) = \sum_{n \in \mathbb{Z}} \Delta(n) \cdot W(n-t) \cdot \exp\{-2\pi j \omega n\}. \quad (5)$$

Note that the frequency parameter ω (measured in Hertz) corresponds to the tempo parameter $\tau = 60 \cdot \omega$ (measured in BPM). Therefore, one obtains a discrete *Fourier tempogram* $\mathcal{T}^{\mathrm{F}} : \mathbb{Z} \times \Theta \to \mathbb{R}_{\geq 0}$ by

$$\mathcal{T}^{\mathrm{F}}(t,\tau) = |\mathcal{F}(t,\tau/60)|. \quad (6)$$

As an example, Fig. 8(a) shows the tempogram $\mathcal{T}^{\mathrm{F}}$ of our Shostakovich example from Fig. 7. Note that $\mathcal{T}^{\mathrm{F}}$ reveals a slightly increasing tempo over time starting with roughly $\tau = 225$ BPM. Also, $\mathcal{T}^{\mathrm{F}}$ reveals the second tempo harmonics starting with $\tau = 450$ BPM. Actually, since the novelty curve Δ locally behaves like a track of positive clicks, it is not hard to see that Fourier analysis responds to harmonics but tends to suppress subharmonics, see also [73], [74].

Also autocorrelation-based methods are widely used to estimate local periodicities [66]. Since these methods, as it turns out, respond to subharmonics while suppressing harmonics, they ideally complement Fourier-based methods, see [73], [74]. To obtain a discrete *autocorrelation tempogram*, one again fixes a window function $W : \mathbb{Z} \to \mathbb{R}$ centered at $t = 0$ with support $[-N : N]$, $N \in \mathbb{N}$. The local autocorrelation is then computed by comparing the windowed novelty curve with time shifted copies of itself. Here, we use the unbiased local autocorrelation

$$\mathcal{A}(t,\ell) = \frac{\sum_{n \in \mathbb{Z}} \Delta(n) W(n-t) \Delta(n+\ell) \cdot W(n-t+\ell)}{2N+1-\ell} \quad (7)$$

for time $t \in \mathbb{Z}$ and time lag $\ell \in [0 : N]$. Now, to convert the lag parameter into a tempo parameter, one needs to know the sampling rate. Supposing that each time parameter $t \in \mathbb{Z}$ corresponds to r seconds, then the lag ℓ corresponds to the tempo $\tau = 60/(r \cdot \ell)$ BPM. From this, one obtains the *autocorrelation tempogram* $\mathcal{T}^{\mathrm{A}}$ by

$$\mathcal{T}^{\mathrm{A}}(t,\tau) = \mathcal{A}(t,\ell). \quad (8)$$

for each tempo $\tau = 60/(r \cdot \ell)$, $\ell \in [1 : N]$. Finally, using standard resampling and interpolation techniques applied to the tempo domain, one can derive an autocorrelation tempogram $\mathcal{T}^{\mathrm{A}} : \mathbb{Z} \times \Theta \to \mathbb{R}_{\geq 0}$ that is defined on the same tempo set $Theta$ as the Fourier tempogram $\mathcal{T}^{\mathrm{F}}$. The tempogram $\mathcal{T}^{\mathrm{A}}$ for our Shostakovich example is shown in Fig. 8(b), which clearly indicates the subharmonics. Actually, the parameter $\tau = 75$ is the third subharmonics of $\tau = 225$ and corresponds to the tempo on the measure level.

Assuming a more or less steady tempo, most tempo estimation approaches determine only one *global* tempo value for the entire recording. For example, such a value may be obtained by averaging the tempo values (e.g., using a median filter [53]) obtained from a framewise periodicity analysis. Dealing with music with significant tempo changes, the task of *local* tempo estimation (for each point in time) becomes a much more difficult or even ill-posed problem; see also Fig. 9 for a complex example. Having computed a tempogram, the framewise maximum yields a good indicator of the locally dominating tempo—however, one often has to struggle with confusions of tempo harmonics and subharmonics. Here, tempo estimation can be improved by a combined usage of Fourier and autocorrelation tempograms. Furthermore, instead of simply taking the framewise maximum, global optimization techniques based on dynamic programming have been suggested to obtain smooth tempo trajectories [61], [67].

C. Beat Tracking

When listening to a piece of music, most humans are able to tap to the musical beat without difficulty. However, transferring this cognitive process into an automated system that reliably works for the large variety of musical styles is a challenging task. In particular, the tracking of beat positions becomes hard in the case that a music recording reveals significant tempo changes. This typically occurs in expressive performances of classical music as a result of ritardandi, accelerandi, fermatas, and artistic shaping [75]. Furthermore, the extraction problem is complicated by the fact that there are various levels that are presumed to contribute to the human perception of beat.

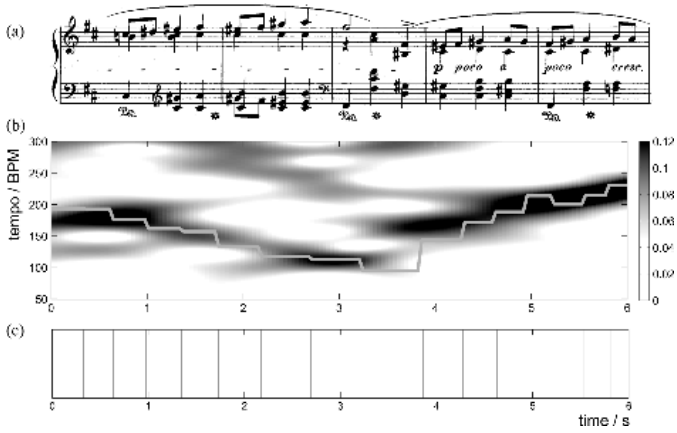

Fig. 9. Excerpt of the Mazurka Op. 30 No. 2 (played by Rubinstein, 1966). (a) Score. (b) Fourier tempogram with reference tempo (cyan). (c) Beat positions (quarter note level).

Most approaches focus on determining musical pulses on the *tactus* level (the foot tapping rate) [65]–[67], but only few approaches exist for analyzing the signal on the measure level [54], [57] or finer tatum level [76]–[78]. Here, a *tatum* or *temporal atom* refers to the fastest repetition rate of musically meaningful accents occurring in the signal [79]. Various approaches have been suggested that simultaneously analyze different pulse levels [57], [68], [80]. In [62] and [63], instead of looking at a specific pulse level, a robust mid-level representation has been introduced which captures the predominant local pulse even in the presence of significant tempo fluctuations.

Exemplarily, we describe a robust beat tracking procedure [66], which assumes a roughly constant tempo throughout the music recording. The input of the algorithm consists of a novelty curve $\Delta : [0 : T-1] \to \mathbb{R}$ as well as an estimate $\hat{\tau}$ of the global (average) tempo, which also determines the pulse level to be considered. From $\hat{\tau}$ and the sampling rate used for the novelty curve, one can derive an estimate $\hat{\rho} \in \mathbb{Z}$ for the average beat period (given in samples). Assuming a roughly constant tempo, the difference δ of two neighboring beats should be close to $\hat{\rho}$. To measure the distance between δ and $\hat{\rho}$, a neighborhood function $N_{\hat{\rho}} : \mathbb{N} \to \mathbb{R}$, $N_{\hat{\rho}}(\delta) = -(\log_2(\delta/\hat{\rho}))^2$, is introduced. This function takes the maximum value of 0 for $\delta = \hat{\rho}$ and is symmetric on a log-time axis. Now, the task is to estimate a sequence $B = (b_1, b_2, \ldots, b_L)$, for some suitable $L \in \mathbb{N}$, of monotonously increasing beat positions $b_\ell \in [0 : T-1]$ satisfying two conditions. On the one hand, the value $\Delta(b_\ell)$ should be large for all $\ell \in [1 : L]$, and, on the other hand, the beat intervals $\delta = b_\ell - b_{\ell-1}$ should be close to $\hat{\rho}$. To this end, one defines the score $S(B)$ of a beat sequence $B = (b_1, b_2, \ldots, b_L)$ by

$$S(B) = \sum_{\ell=1}^{L} \Delta(b_\ell) + \alpha \sum_{\ell=2}^{L} N_{\hat{\rho}}(b_\ell - b_{\ell-1}) \qquad (9)$$

where the weight $\alpha \in \mathbb{R}$ balances out the two conditions. Finally, the beat sequence maximizing S yields the solution of the beat tracking problem. The score-maximizing beat sequence can be obtained by a straightforward dynamic programming (DP) approach; see [66] for details.

As mentioned above, recent beat tracking procedures work well for modern pop and rock music with a strong and steady beat, but the extraction of beat locations from highly expressive performances still constitutes a challenging task with many open problems. For such music, one often has significant local tempo fluctuation caused by the artistic freedom a musician takes, so that the model assumption of local periodicity is strongly violated. This is illustrated by Fig. 9, which shows a tempo curve and the beat positions for a romantic piano music recording (Mazurka by Chopin). In practice beat tracking is further complicated by the fact that there may be beats with no explicit note events going along with them [81]. Here, a human may still perceive a steady beat by subconsciously interpolating the missing onsets. This is a hard task for a machine, in particular in passages of varying tempo where interpolation is not straightforward. Furthermore, auxiliary note onsets can cause difficulty or ambiguity in defining a specific physical beat time. In music such as the Chopin Mazurkas, the main melody is often embellished by ornamented notes such as trills, grace notes, or arpeggios. Also, for the sake of expressiveness, the notes of a chord need not be played at the same time, but slightly displaced in time. This renders a precise definition of a physical beat position impossible [82]. Such highly expressive music also reveals the limits of purely onset-oriented tempo and beat tracking procedures, see also [63].

D. Higher-Level Rhythmic Structures

The extraction of onset, beat, and tempo information is of fundamental importance for the determination of higher-level musical structures such as rhythm and meter [46], [48]. Generally, the term *rhythm* is used to refer to a temporal patterning of event durations, which are determined by a regular succession of strong and weak stimuli [83]. Furthermore, the perception of rhythmic patterns also depends on other cues such as the dynamics and timbre of the involved sound events. Such repeating patterns of accents form characteristic *pulse groups*, which determine the *meter* of a piece of music. Here, each group typically starts with an accented beat and consists of all pulses until the next accent. In this sense, the term *meter* is often used synonymously with the term *time signature*, which specifies the beat structure of a musical measure or bar. It expresses a regular pattern of beat stresses continuing through a piece thus defining a hierarchical grid of beats at various time scales.

Rhythm and tempo are often sufficient for characterizing the style of a piece of music. This particularly holds for dance music, where, e.g., a waltz or tango can be instantly recognized from the underlying rhythmic pattern. Various approaches have been described for determining some kind of rhythm template, which have mainly been applied for music classification tasks [77], [84], [85]. Typically, the first step consists in performing some beat tracking. In the next step, assuming additional knowledge such as the time signature and the starting position of the first bar, patterns of alternating strong and weak pulses are determined for each bar, which are then averaged over all bars to yield an average rhythmic pattern for the entire piece [84], [85]. Even though such patterns may still be abstract, they have been successfully applied for tasks such as dance style classification. The automatic extraction of explicit rhythmic

parameters such as the time signature constitutes a difficult problem. A first step towards time signature estimation has been described in [86], where the number of beats between regularly recurring accents (or downbeats) are estimated to distinguish between music having a duple or triple meter.

Another way for deriving rhythm-related features is to consider intervals defined by successive onset or beat positions, often referred as *inter-onset-intervals* (IOIs). Considering histograms over the durations of occurring IOIs, one may then derive hypotheses on the beat period, tempo, and meter [75], [76], [87]. The drawback of these approaches is that they rely on an explicit localization of a discrete set of onset and beat positions—a fragile and error-prone step. To compensate for such errors, various approaches have been proposed to jointly or iteratively estimate onset, pulse, and meter parameters [54], [78].

IV. Timbre and Instrumentation

Timbre is defined as the "attribute of auditory sensation in terms of which a listener can judge two sounds similarly presented and having the same loudness and pitch as dissimilar" [88]. The concept is closely related to sound source recognition: for example, the sounds of the violin and the flute may be identical in their pitch and loudness, but are still easily distinguished. Furthermore, when listening to polyphonic music, we are usually able to perceptually organize the component sounds to their sources based on timbre information.

The term *polyphonic timbre* refers to the overall timbral mixture of a music signal, the "global sound" of a piece of music [89], [90]. Human listeners, especially trained musicians, can switch between a "holistic" listening mode where they consider a music signal as a coherent whole, and a more analytic mode where they focus on the part played by a particular instrument [91], [92]. In computational systems, acoustic features describing the polyphonic timbre have been found effective for tasks such as automatic genre identification [93], music emotion recognition [94], and automatic tagging of audio with semantic descriptors [95]. A computational analogy for the analytical listening mode, in turn, includes recognizing musical instruments on polyphonic recordings.

This section will first discuss feature representations for timbre and then review methods for musical instrument recognition in isolation and in polyphonic music signals.

A. Perceptual Dimensions of Timbre

Timbre is a multidimensional concept, having several underlying acoustic factors. Schouten [96] describes timbre as being determined by five major acoustic parameters: 1) the range between tonal and noise-like character; 2) the spectral envelope; 3) the time envelope; 4) the changes of spectral envelope and fundamental frequency; and 5) the onset of the sound differing notably from the sustained vibration.

The perceptual dimensions of timbre have been studied based on dissimilarity ratings of human listeners for sound pairs; see [97] and [98]. In these studies, multidimensional scaling (MDS) was used to project the dissimilarity ratings into a lower-dimensional space where the distances between the sounds match as closely as possible the dissimilarity ratings. Acoustic correlates can then be proposed for each dimension of this timbre space. Several studies report spectral centroid, $c_t = \sum_k |kX(t,k)| / \sum_k |X(t,k)|$, and attack time as major determinants of timbre. Also often reported are spectral irregularity (defined as the average level difference of neighboring harmonics) and spectral flux, $d = (1/T)\sum_t ||\mathbf{d}_t||$, where $\mathbf{d}_t = |X(t,k)| - |X(t-1,k)|$.

Very few studies have attempted to uncover the perceptual dimensions of polyphonic timbre. Cogan [99] carried out informal musicological case studies using the spectrograms of diverse music signals and proposed 13 dimensions to describe the quality of musical sounds. Furthermore, Kendall and Carterette [100] studied the perceptual dimensions of simultaneous wind instrument timbres using MDS, whereas Alluri and Toiviainen [90] explored the polyphonic timbre of Indian popular music. The latter observed relatively high correlations between certain perceptual dimensions and acoustic features describing spectro-temporal modulations.

B. Time-Varying Spectral Envelope

The acoustic features found in the MDS experiments bring insight into timbre perception, but they are generally too low-dimensional to lead to robust musical instrument identification [101]. In signal processing applications, timbre is typically described using a parametric model of the time-varying spectral envelope of sounds. This stems from speech recognition [102] and is not completely satisfactory in music processing as will be seen in Section IV-C, but works well as a first approximation of timbre. Fig. 10 illustrates the time-varying spectral envelopes of two example musical tones. Indeed, all the acoustic features found in the MDS experiments are implicitly represented by the spectral envelope, and among the five points on Schouten's list, 2)–5) are reasonably well covered. The first point, tonal versus noiselike character, can be addressed by decomposing a music signal into its sinusoidal and stochastic components [103], [104], and then estimating the spectral envelope of each part separately. This, for example, has been found to significantly improve the accuracy of genre classification [105].

Mel-frequency cepstral coefficients (MFCCs), originally used for speech recognition [102], are by far the most popular way of describing the spectral envelope within an individual analysis frame. MFCCs encode the coarse shape of the log-power spectrum on the mel-frequency scale.[4] They have the desirable property that a small (resp. large) numerical change in the MFCC coefficients corresponds to a small (resp. large) perceptual change. MFCCs are calculated by simulating a bank of about 40 bandpass filters in the frequency domain (the filters being uniformly spaced on the Mel-frequency scale), calculating the log-power of the signal within each band, and finally applying a discrete cosine transform to the vector of log-powers to obtain the MFCC coefficients. Typically only the 10–15 lowest coefficients are retained and the rest are discarded in order to make the timbre features invariant to pitch information that is present in the higher coefficients. Time-varying aspects are usually accounted for by appending temporal derivatives of the MFCCs to the feature vector.

[4]The mel-frequency scale is one among several scales that model the frequency resolution of the human auditory system. See [106] for a comparison of different perceptual frequency scales.

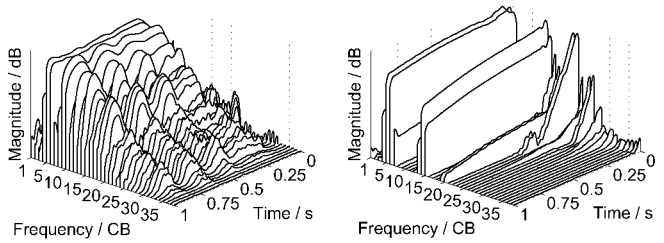

Fig. 10. Time-varying spectral envelopes of 260-Hz tones of the flute (left) and the vibraphone (right). Here sound pressure level within auditory critical bands (CB) is shown as a function of time.

Modulation spectrum encodes the temporal variation of spectral energy explicitly [107]. This representation is obtained by first using a filterbank to decompose an audio signal into subbands, extracting the energy envelope within each band, and finally analyzing amplitude modulations (AM) within each band by computing discrete Fourier transforms of the energy envelope within longer "texture" windows (phases are discarded to achieve shift-invariance). This results in a three-dimensional representation where the dimensions correspond to time, frequency, and AM frequency (typically in the range 0–200 Hz). Sometimes the time dimension can be collapsed by analyzing AM modulation in a single texture window covering the entire signal. Spectro-temporal modulations play an important role in the perception of polyphonic timbre [90]; therefore, representations based on modulation spectra are particularly suitable for describing the instrumentation aspects of complex music signals. Indeed, state-of-the-art genre classification is based on the modulation spectrum [108]. Other applications of the modulation spectrum include speech recognition [109], audio coding [107], and musical instrument recognition [110].

C. Source-Filter Model of Sound Production

Let us now consider more structured models of musical timbre. Instrument acoustics provides a rich source of information for constructing models for the purpose of instrument recognition. The source-filter model of sound production is particularly relevant here [111]. Many musical instruments can be viewed as a coupling of a vibrating object, such as a guitar string ("source"), with the resonance structure of the rest of the instrument ("filter") that colors the produced sound. The source part usually determines pitch, but often contains also timbral information.

The source-filter model has been successfully used in speech processing for decades [112]. However, an important difference between speech and music is that there is only one sound production mechanism in speech, whereas in music a wide variety of sound production mechanisms are employed. Depending on the instrument, the sound can be produced for example by vibrating strings, air columns, or vibrating bars, and therefore the source excitation provides valuable information about the instrument identity.

It is interesting to note that the regularities in the source excitation are not best described in terms of frequency, but in terms of harmonic index. For example the sound of the clarinet is characterized by the odd harmonics being stronger than the even harmonics. For the piano, every Mth partial is weaker because the string is excited at a point $1/M$ along its length. The sound of the vibraphone, in turn, exhibits mainly the first and the fourth harmonic and some energy around the tenth partial. MFCCs and other models that describe the properties of an instrument as a function of frequency smear out this information. Instead, a structured model is needed where the spectral information is described both as a function of frequency and as a function of harmonic index.

The source-filter model for the magnitude spectrum $|X(f)|$ of a harmonic sound can be written as

$$|X(f_h)| = \gamma S(h) B(f_h) \tag{10}$$

where $f_h \approx hv$, $h = 1, 2, \ldots$ is the frequency of the hth harmonic of a sound with fundamental frequency v. Note that $|X(f)|$ is modeled only at the positions of the harmonics and is assumed zero elsewhere. The scalar γ denotes the overall gain of the sound, $S(h)$ is the amplitude of harmonic h in the spectrum of the vibrating source, and $B(f)$ represents the frequency response of the instrument body. Perceptually, it makes sense to minimize the modeling error on the log-magnitude scale and therefore to take the logarithm of both sides of (10). This renders the model linear and allows the two parts, $\log S(h)$ and $\log B(f_h)$, to be further represented using a suitable linear basis [113]. In addition to speech coding and music synthesis [111], [112], the source-filter model has been used to separate the main melody from polyphonic music [114] and to recognize instruments in polyphonic music [115].

Above we assumed that the source excitation produces a spectrum where partial frequencies obey $f_h \approx hv$. Although such sounds are the commonplace in Western music (mallet percussion instruments being the exception), this is not the case in all music cultures and the effect of partial frequencies on timbre has been very little studied. Sethares [116] has investigated the relationship between the spectral structure of musical sounds and the structure of musical scales used in different cultures.

D. Recognition of Musical Instruments in Isolation

This section reviews techniques for musical instrument recognition in signals where only one instrument is playing at a time. Systems developed for this purpose typically employ the supervised classification paradigm (see Fig. 11), where 1) acoustic features are extracted in successive time frames in order to describe the relevant aspects of the signal; 2) training data representing each instrument class is used to learn a model for within-class feature distributions; and 3) the models are then used to classify previously unseen samples.

A number of different supervised classification methods have been used for instrument recognition, including k-nearest neighbors, Gaussian mixture models, hidden Markov models, linear discriminant analysis, artificial neural networks, support vector machines, and decision trees [117], [118]. For a comprehensive review of the different recognition systems for isolated notes and solo phrases, see [119]–[122].

A variety of acoustic features have been used for instrument recognition. Spectral features include the first few moments of the magnitude spectrum (spectral centroid, spread, skewness,

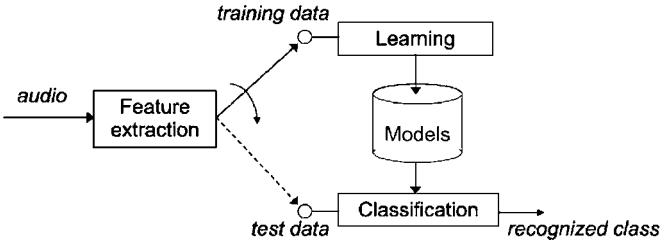

Fig. 11. General overview of supervised classification. See text for details.

and kurtosis), sub-band energies, spectral flux, spectral irregularity, and harmonic versus noise part energy [123]. Cepstral features include MFCCs and warped linear prediction-based cepstral coefficients, see [101] for a comparison. Modulation spectra have been used in [110]. Temporal features include the first few moments of the energy envelope within frames of about one second in length, and the frequency and strength of amplitude modulation in the range 4–8 Hz ("tremolo") and 10–40 Hz ("roughness") [122], [124]. The first and second temporal derivatives of the features are often appended to the features vector. For a more comprehensive list of acoustic features and comparative evaluations, see [101], [122], [125], [126].

Obviously, the above list of acoustic features is highly redundant. Development of a classification system typically involves a feature selection stage, where training data is used to identify and discard unnecessary features and thereby reduce the computational load of the feature extraction, see Herrera *et al.* [119] for a discussion on feature selection methods. In order to facilitate the subsequent statistical modeling of the feature distributions, the retained features are often decorrelated and the dimensionality of the feature vector is reduced using principal component analysis, linear discriminant analysis, or independent component analysis [117].

Most instrument classification systems resort to the so-called bag-of-features approach where an audio signal is modeled by the statistical distribution of its short-term acoustic features, and the temporal order of the features is ignored. An exception here are the instrument recognizers employing hidden Markov models where temporal dependencies are taken into account explicitly [127], [128]. Joder *et al.* [122] carried out an extensive evaluation of different temporal integration mechanisms to see if they improve over the bag-of-features approach. They found that a combination of feature-level and classifier-level temporal integration improved over a baseline system, although neither of them alone brought a significant advantage. Furthermore, HMMs performed better than GMMs, which suggests that taking into account the temporal dependencies of the feature vectors improves classification.

The techniques discussed above are directly applicable to other audio signal classification tasks too, including genre classification [93], automatic tagging of audio [95], and music emotion recognition [94], for example. However, the optimal acoustic features and models are usually specific to each task.

E. Instrument Recognition in Polyphonic Mixtures

Instrument recognition in polyphonic music is closely related to sound source separation: recognizing instruments in a mixture allows one to generate time–frequency masks that indicate which spectral components belong to which instrument. Vice versa, if individual instruments can be reliably separated from the mixture, the problem reduces to that of single-instrument recognition. The problem of source separation will be discussed in Section V.

A number of different approaches have been proposed for recognizing instruments in polyphonic music. These include extracting acoustic features directly from the mixture signal, sound source separation followed by the classification of each separated signal, signal model-based probabilistic inference, and dictionary-based methods. Each of these will be discussed in the following.

The most straightforward approach to polyphonic instrument recognition is to extract features directly from the mixture signal. Little and Pardo [129] used binary classifiers to detect the presence of individual instruments in polyphonic audio. They trained classifiers using weakly labeled mixture signals, meaning that only the presence or absence of the target sound object was indicated but not the exact times when it was active. They found that learning from weakly labeled mixtures led to better results than training with isolated examples of the target instrument. This was interpreted to be due to the fact that the training data, in the mixed case, was more representative of the polyphonic data on which the system was tested. Essid *et al.* [124] developed a system for recognizing *combinations* of instruments directly. Their method exploits hierarchical classification and an automatically built taxonomy of musical ensembles in order to represent every possible combination of instruments that is likely to be played simultaneously in a given genre.

Eggink and Brown [130] introduced missing feature theory to instrument recognition. Here the idea is to estimate a binary mask that indicates time–frequency regions that are dominated by energy from interfering sounds and are therefore to be excluded from the classification process [131]. The technique is known to be effective if the mask is correctly estimated, but estimating it automatically is hard. Indeed, Wang [132] has proposed that estimation of the time–frequency masks of sound sources can be viewed as the computational goal of auditory scene analysis in general. Fig. 12 illustrates the use of binary masks in the case of a mixture consisting of singing and piano accompaniment. Estimating the mask in music is complicated by the fact that consonant pitch intervals cause partials of different sources to co-incide in frequency. Kitahara *et al.* [133] avoided the mask estimation by applying linear discriminant analysis on features extracted from polyphonic training data. As a result, they obtained feature weightings where the largest weights were given to features that were least affected by the overlapping partials of co-occurring sounds.

A number of systems are based on separating the sound sources from a mixture and then recognizing each of them individually [115], [123], [134]–[136]. Heittola *et al.* [115] used a source-filter model for separating the signals of individual instruments from a mixture. They employed a multiple-F0 estimator to produce candidate F0s at each time instant, and then developed a variant of the non-negative matrix factorization algorithm to assign sounds to their respective instruments

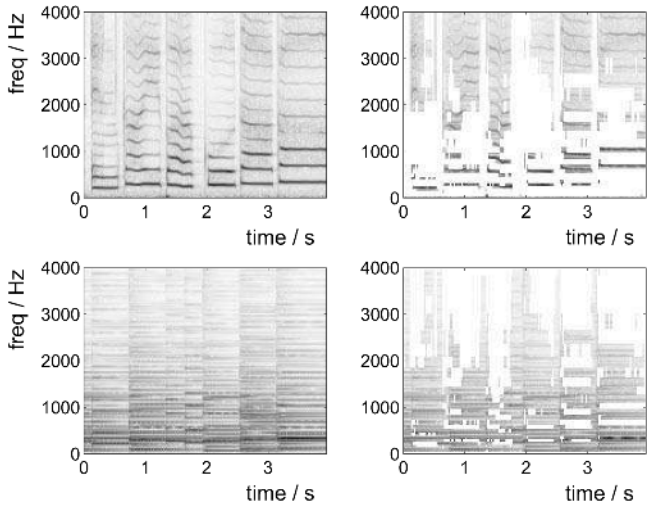

Fig. 12. Illustration of the use of binary masks. The left panels show the magnitude spectrograms of a singing excerpt (top) and its piano accompaniment (bottom). The right panels show spectrograms of the mixture of the singing and the accompaniment, with two different binary masks applied. On the top-right, white areas indicate regions where the accompaniment energy is higher than that of singing. On the bottom-right, singing has been similarly masked out.

and to estimate the spectral envelope of each instrument. A different approach to sound separation was taken by Martins *et al.* [135] and Burred *et al.* [136] who employed ideas from computational auditory scene analysis [137]. They extracted sinusoidal components from the mixture spectrum and then used cues such as common onset time, frequency proximity, and harmonic frequency relationships to assign spectral components to distinct groups. Each group of sinusoidal trajectories was then sent to a recognizer.

Vincent and Rodet [138] viewed instrument recognition as a parameter estimation problem for a given signal model. They represented the short-term log-power spectrum of polyphonic music as a weighted nonlinear combination of typical note spectra plus background noise. The note spectra for each instrument were learnt in advance from a database of isolated notes. Parameter estimation was carried out by maximizing the joint posterior probability of instrument labels and the activation parameters E_{ijt} of note j and instrument i at time t. Maximizing this joint posterior resulted in joint instrument recognition and polyphonic transcription.

A somewhat different path can also be followed by relying on sparse decompositions. The idea of these is to represent a given signal with a small number of elements drawn from a large (typically overcomplete) dictionary. For example, Leveau *et al.* [139] represented a time-domain signal $x(t)$ as a weighted sum of atoms $h_\lambda(t)$ taken from a dictionary $\mathcal{D} = \{h_\lambda(t)\}_\lambda$, and a residual $r(t)$:

$$x(t) = \sum_{\lambda \in \Lambda} a_\lambda h_\lambda(t) + r(t) \tag{11}$$

where Λ is a finite set of indexes λ. Each atom $h_\lambda(t)$ consists of a sum of M windowed and amplitude-weighted sinusoidals at frequencies that are integer multiples of a linearly varying fundamental frequency. An individual atom covers only a short frame of the input signal, but continuity constraints can be placed on the activations a_λ of atoms with successive temporal supports. Leveau *et al.* [139] learned the dictionary of atoms in advance from a database of isolated musical tones. A sparse decomposition for a given mixture signal was then found by maximizing the signal-to-residual ratio for a given number of atoms. This optimization process results in selecting the most suitable atoms from the dictionary, and since the atoms have been labeled with pitch and instrument information, this results in joint instrument identification and polyphonic transcription. Also the instrument recognition methods of Kashino and Murase [140] and Cont and Dubnov [110] can be viewed as being based on dictionaries, the former using time-domain waveform templates and the latter modulation spectra.

The above-discussed sparse decompositions can be viewed as a mid-level representation, where information about the signal content is already visible, but no detection or thresholding has yet taken place. Such a goal was pursued by Kitahara *et al.* [128] who proposed a "note-estimation-free" instrument recognition system for polyphonic music. Their system used a spectrogram-like representation ("instrogram"), where the two dimensions corresponded to time and pitch, and each entry represented the probability that a given target instrument is active at that point.

V. Polyphony and Musical Voices

Given the extensive literature of speech signal analysis, it seems natural that numerous music signal processing studies have focused on monophonic signals. While monophonic signals certainly result in better performance, the desire for wider applicability has led to a gradual focus, in recent years, to the more challenging and more realistic case of polyphonic music. There are two main strategies for dealing with polyphony: the signal can either be processed globally, directly extracting information from the polyphonic signal, or the system can attempt to first split up the signal into individual components (or sources) that can then be individually processed as monophonic signals. The source separation step of this latter strategy, however, is not always explicit and can merely provide a mid-level representation that facilitates the subsequent processing stages. In the following sections, we present some basic material on source separation and then illustrate the different strategies on a selection of specific music signal processing tasks. In particular, we address the tasks of multi-pitch estimation and musical voice extraction including melody, bass, and drum separation.

A. Source Separation

The goal of source separation is to extract all individual sources from a mixed signal. In a musical context, this translates in obtaining the individual track of each instrument (or individual notes for polyphonic instruments such as piano). A number of excellent overviews of source separation principles are available; see [141] and [142].

In general, source separation refers to the extraction of full bandwidth source signals but it is interesting to mention that several polyphonic music processing systems rely on a simplified source separation paradigm. For example, a filter bank decomposition (splitting the signal in adjacent well defined frequency bands) or a mere Harmonic/Noise separation [143] (as for drum extraction [144] or tempo estimation [61]) may be regarded as instances of rudimentary source separation.

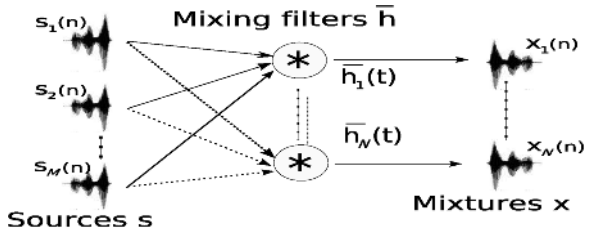

Fig. 13. Convolutive mixing model. Each mixture signal $\mathbf{x}_i(n)$ is then expressed from the source signals as: $\mathbf{x}_i(n) = \sum_{j-1}^{M} h_{ij}(t) * \mathbf{s}_j(n)$.

Three main situations occur in source separation problems. The *determined case* corresponds to the situation where there are as many mixture signals as different sources in the mixtures. Contrary, the *overdetermined* (resp. *underdetermined*) case refers to the situation where there are more (resp. less) mixtures than sources. Underdetermined Source Separation (USS) is obviously the most difficult case. The problem of source separation classically includes two major steps that can be realized jointly: estimating the mixing matrix and estimating the sources. Let $\mathbf{X} = [\mathbf{x}_1(n), \ldots, \mathbf{x}_N(n)]^T$ be the N mixture signals, $\mathbf{S} = [\mathbf{s}_1(n), \ldots, \mathbf{s}_M(n)]^T$ the M source signals, and $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N]^T$ the $N \times M$ mixing matrix with mixing gains $\mathbf{a}_i = (a_{i1}, a_{i2}, \ldots, a_{iM})$. The mixture signals are then obtained by: $\mathbf{X} = \mathbf{AS}$. This readily corresponds to the *instantaneous mixing* model (the mixing coefficients are simple scalars). The more general convolutive mixing model considers that a filtering occurred between each source and each mixture (see Fig. 13). In this case, if the filters are represented as $N \times M$ FIR filters of impulse response $h_{ij}(t)$, the mixing matrix is given by $\mathbf{A} = (\overline{\mathbf{h}}_1(t), \overline{\mathbf{h}}_2(t), \ldots, \overline{\mathbf{h}}_N(t))^T$ with $\overline{\mathbf{h}}_i(t) = [h_{i1}(t), \ldots, h_{iM}(t)]$, and the mixing model corresponds to $\mathbf{X} = \mathbf{A} * \mathbf{S}$.

A wide variety of approaches exist to estimate the mixing matrix and rely on techniques such as Independent Component Analysis (ICA), sparse decompositions or clustering approaches [141]. In the determined case, it is straightforward to obtain the individual sources once the mixing matrix is known: $\mathbf{S} = \mathbf{A}^{-1}\mathbf{X}$. The underdetermined case is much harder since it is an ill-posed problem with an infinite number of solutions. Again, a large variety of strategies exists to recover the sources including heuristic methods, minimization criteria on the error $||\mathbf{X} - \mathbf{AS}||^2$, or time–frequency masking approaches. One of the popular approaches, termed adaptive Wiener filtering, exploits soft time–frequency masking. Because of its importance for audio source separation, it is described in more details.

For the sake of clarity, we consider below the monophonic case, i.e., where only one mixture signal $\mathbf{x}_1(n)$ is available. If we consider that the M sources $\mathbf{s}_i(n)$ are stationary Gaussian processes of power spectral density (PSD) $\sigma_i^2(k)$, then the optimal estimate of $\mathbf{s}_i(n)$ is obtained as

$$S_i(n,k) = \frac{\sigma_i^2(k)}{\sum_{j=1}^{M} \sigma_j^2(k)} X_1(n,k) \tag{12}$$

where $X_1(n,k)$ and $S_i(n,k)$ are the STFTs of the mixture $\mathbf{x}_1(n)$ and source $\mathbf{s}_i(n)$, respectively. In practice, audio signals can only be considered as locally stationary and are generally assumed to be a combination of stationary Gaussian processes. The source signal $\mathbf{s}_i(n)$ is then given by

$$\mathbf{s}_i(n) = \sum_{m \in K_i} \alpha_m(n) b_m(n),$$

where $b_m(n)$ are stationary Gaussian processes of PSD $\sigma_m^2(k)$, $\alpha_m(n) \geq 0$ are slowly varying coefficients, and K_i is a set of indices for source $\mathbf{s}_i(n)$. Here, the estimate of $\mathbf{s}_i(n)$ is then obtained as (see for example [145] or [144] for more details):

$$S_i(n,k) = \frac{\sum_{m \in K_i} \alpha_m(n) \sigma_m^2(k)}{\sum_{m \in K_1 \cup \ldots \cup K_M} \alpha_m(n) \sigma_m^2(k)} X_1(n,k). \tag{13}$$

Note that in this case, it is possible to use decomposition methods on the mixture $\mathbf{x_1}(n)$ such as non-negative matrix factorization (NMF) to obtain estimates of the spectral templates $\sigma_m^2(k)$.

Music signal separation is a particularly difficult example of USS of convolutive mixtures (many concurrent instruments, possibly mixed down with different reverberation settings, many simultaneous musical notes and, in general, a recording limited to two channels). The problem is then often tackled by integrating prior information on the different source signals. For music signals, different kinds of prior information have been used including timbre models [146], harmonicity of the sources [147], temporal continuity, and sparsity constraints [148]. In some cases, by analogy with speech signal separation, it is possible to exploit production models, see [114] or [149].

Concerning evaluation, the domain of source separation of audio signals is also now quite mature and regular evaluation campaigns exist[5] along with widely used evaluation protocols [150].

B. From Monopitch to Multipitch Estimation

The estimation of the fundamental frequency of a quasi-periodic signal, termed *monopitch estimation*, has interested the research community for decades. One of the main challenges is to obtain a versatile and efficient algorithm for a wide range of possible fundamental frequencies which can cope with the deviations of real audio signals from perfect periodicity. For speech signals, extensive reviews of early algorithms can be found in [151] and [152].

In general, the different methods can be roughly classified in three classes depending on the used signal representation. The *frequency domain approaches* exploit the fact that quasi-periodic signals exhibit a quasi-harmonic distribution of peaks in the spectral domain. The fundamental frequency is estimated by searching the highest frequency that generates a spectral comb best explaining the spectral content of the signal; see [153]–[155]. The *time domain approaches* aim at directly estimating the period on the basis of the signal's waveform by searching the smallest time-shift for which the waveform and its time-shifted version match. This can be done using the autocorrelation or the average magnitude difference functions [156], [157] or applying kernel-based approaches [158]. Both time and frequency domain approaches are prone to octave

[5]For example, see http://sisec.wiki.irisa.fr/tiki-index.php (SiSec campaign).

errors. However, because frequency domain (resp. time-domain) approaches are prone to estimate integer multiples or harmonics (resp. integer fractions or sub-harmonics) of the true fundamental frequency, *mixed domain approaches* exploiting both representations were also developed; see [159] or [160].

Even though monopitch estimation is now achievable with a reasonably high accuracy, the problem of multipitch estimation (e.g., estimating the fundamental frequency of concurrent periodic sounds) remains very challenging. The problem is indeed particularly difficult for music signals for which concurrent notes stand in close harmonic relation. Here, in some sense, the worst case is when two simultaneous notes are played one or several octaves apart. For extreme cases such as complex orchestral music, where one has a high level of polyphony, multipitch estimation becomes intractable with today's methods. For a review of recent approaches, we refer to [161]–[163]. Most approaches work, at least partially, in the spectral domain. On the one hand, some methods follow a global strategy aiming at jointly estimating all fundamental frequencies. For example, [164] or [165] employ parametric methods, [166] describes a dedicated comb approach, [167] and [168] use methods based on machine learning paradigms, whereas [169] follows a least-square strategy. On the other hand, a number of methods rely on source separation principles which are more or less explicit. For example, the non-negative matrix factorization framework was successfully used in several algorithms [147], [148], [170]. In other approaches, the source separation may be less explicit and can, for example, rely on an iterative procedure [171]. Here, the dominant fundamental frequency is first estimated, then the spectral peaks of the corresponding musical note are identified and subtracted (sometimes only partially) from the polyphonic signal to obtain a residual signal. The procedure is iterated while the residual contains at least one musical note. Despite the inherent limitations of iterative procedures, these approaches are among the most efficient to date as regularly shown in the MIREX evaluation campaign.[6]

It is worth emphasizing that most methods exploit musical knowledge in one or the other way. Characteristic timbre information may be used in the form of instrument spectral models or templates as prior information to better separate the musical sources (as in [172]). Also, spectral smoothness principles can be exploited to subtract more realistic musical notes in iterative methods (as in [171]). Furthermore, constraints on the synchronous evolution of partials amplitude may further help to identify the spectral contribution of a note in a mixture [165]. The specificities of the production mechanism can also be exploited, e.g., in the form of a source/filter production model for the separation, or the introduction of inharmonicities in the model for instruments such as piano [173]. It is also possible to reach higher performances by means of duration or note evolution models (for example based on hidden Markov models as in [173], [174]). This, indeed, permits to take advantage of observations in successive time frames. Finally, knowledge of auditory perception has also been used with success in a number of methods mostly as a front-end acoustic analysis, see [159] and [175]. It is believed that future progress will be fueled by a better understanding of the perception of sound mixtures. Here, an interesting question is why a trained musician has no problem in analyzing a chord containing two notes one octave apart. Better understanding in which way two overlapping partials interact and how their amplitude can be precisely estimated is certainly of central importance [176].

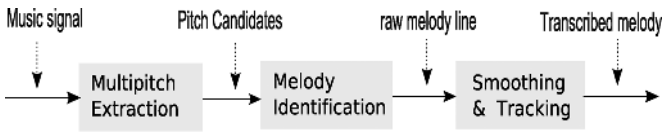

Fig. 14. General scheme underlying most melody transcription systems. From [183].

C. Main Melody or Singing Voice Extraction

Main melody extraction, especially for the singing voice, has received a great deal of interest. This is mainly motivated by the wide range of potential applications including karaoke [177], query-by-humming [178], lead-sheet generation [179], query-by-examples and cover version detection [12], [33], [180]. Following the definition by Paiva [181], "*Melody is the dominant individual pitched line in a musical ensemble*," the task is often restricted to a mere predominant-F0 estimation and tracking task [182]. Only few studies address the full problem leading to a musical score of the melody line, which integrates a note segmentation stage [174], [179]. As described in [183], the problem of main melody extraction is traditionally split into a preliminary analysis stage followed by a melody identification phase and concluded by a smoothing or tracking process; see Fig. 14.

The analysis stage can directly output a raw sequence of predominant fundamental frequency candidates (as in [174] or [181]) but can also produces an intermediate representation or probabilistic model with posterior probabilities for each potential note that would be further exploited in the melody tracking stage (as in [182] or [114]). The analysis stage mostly relies on a spectral domain representation as obtained by a traditional short-time Fourier transform, but may also be obtained by specific multi-resolution transforms [184] or perceptually motivated representations [181].

In the melody identification phase, the sequence of fundamental frequencies that most likely corresponds to the melody is identified using, for example, ad hoc rules, constrained dynamic programming, or hidden Markov models. Finally, if not integrated in the previous stage, a final smoothing or tracking process occurs where the initial estimation of the melody line is further smoothed in order to avoid sudden jumps in the melody line. Such jumps may be caused by initial octave errors or other extraction errors.

In terms of performance, it appears that the most accurate algorithm evaluated in the MIREX evaluation campaigns follows a rule based approach [185] although statistical based systems indicate very promising directions for the future.[7]

To provide some insight on main melody extraction, we now describe one of the existing statistical approaches in more detail [114], [186]. For the sake of clarity, we consider here the

[6]See http://www.music-ir.org/mirex/wiki/.

[7]www.music-ir.org/mirex/wiki/2009:Audio_Melody_Extraction_Results.

monophonic case, where a single mixture signal $\mathbf{x}_1(n)$ is observed. In this model, the observed signal $\mathbf{x}_1(n)$ is the sum of two contributions: the leading voice $\mathbf{v}(n)$ and the musical background $\mathbf{m}(n)$. For a given frame, the STFT of the mixture signal $X(n,k)$ can then be expressed as $X(n,k) = V(n,k) + M(n,k)$, where $V(n,k)$ and $M(n,k)$ are the STFTs of $\mathbf{v}(n)$ and $\mathbf{m}(n)$, respectively. Furthermore, $V(n,k)$ and $M(n,k)$ are assumed to be center proper Gaussians:[8]

$$V(n,k) \sim \mathcal{N}_c\left(0, \mathrm{diag}\left(\sigma^2_{n,V}(k)\right)\right) \quad (14)$$

$$M(n,k) \sim \mathcal{N}_c\left(0, \mathrm{diag}\left(\sigma^2_{n,M}(k)\right)\right) \quad (15)$$

where $\sigma^2_{n,V}(k)$ (resp. $\sigma^2_{n,M}(k)$) is the power spectral density (PSD) of the leading voice $\mathbf{v}(n)$ (resp. of the background music $\mathbf{m}(n)$). Assuming that the leading voice and the background music are independent, the mixture signal at frame n is also a proper Gaussian vector:

$$X(n,k) \sim \mathcal{N}_c\left(0, \mathrm{diag}\left(\sigma^2_{n,X}(k)\right)\right) \quad (16)$$

$$\sim \mathcal{N}_c\left(0, \mathrm{diag}\left(\sigma^2_{n,V}(k) + \sigma^2_{n,M}(k)\right)\right). \quad (17)$$

For extracting the main melody, one then needs to estimate $\sigma^2_{n,V}(k)$ and $\sigma^2_{n,M}(k)$. This can be done by expressing the PSDs as Gaussian mixture models learned on dedicated databases [145]. In [114], assuming a singing voice, the approach is entirely unsupervised, i.e., no learning step is involved. Instead it relies on specific constraints for the voice and the musical background signal. More precisely, the voice signal is assumed to follow a source/filter production model where the source is a periodic signal (referring to the periodic glottal pulse of the singing voice) and where the filter is constrained to smoothly evolve (referring to the slowly varying vocal tract shapes while singing). For the musical background signal, no specific constraints are assumed because of the wide variability of possible music instruments. The estimation of the various model parameters is then conducted by iterative approaches based on NMF techniques. Once the PSDs $\sigma^2_{n,V}(k)$ and $\sigma^2_{n,M}(k)$ of both signals are obtained, the separated singing voice signal is obtained using the Wiener filter approach for each frame as in (12).

Since the model is rather generic, it is also directly applicable to other leading instruments such as a trumpet within a Jazz quartet; see Fig. 15.

D. Bass Line Extraction

Following [174], the term *bass line* refers to an organized sequence of consecutive notes and rests played with a bass guitar, a double bass or a bass synthesizer. The bass line plays an important role in several music styles particularly in popular music. Having a separated bass line or a transcription of the bass line opens the path to a number of applications including "music minus one" for bass players or various indexing tasks such as chord extraction, downbeat estimation, music genre or mood classification [187].

[8]A complex proper Gaussian random variable is a complex random variable whose real part and imaginary part are independent and follow a (real) Gaussian distribution, with the same parameters: mean equal to 0 and identical variance (co-variance matrix in the multi-variate case).

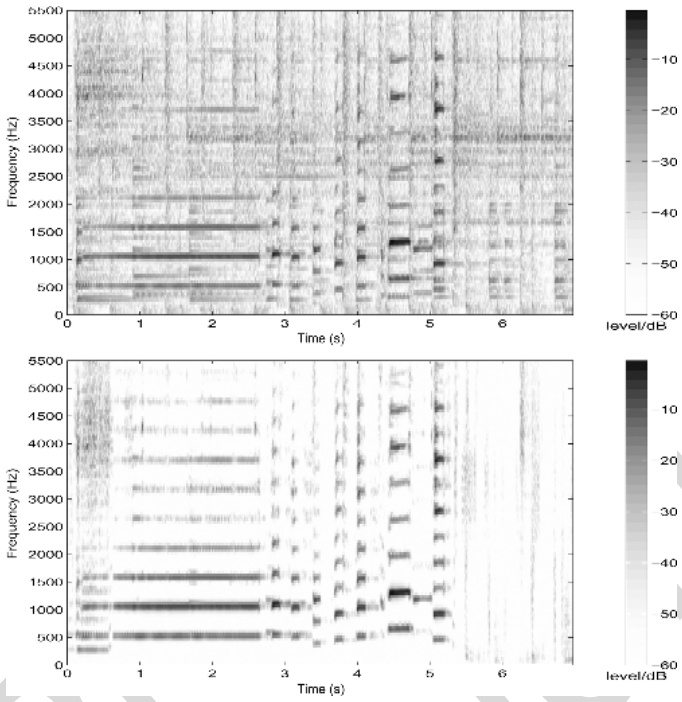

Fig. 15. Spectrograms of the original signal (top) and of the separated trumpet signal (bottom) of the piece *Caravan* played by the *Marsalis Jazz Quartet*. From [114].

Bass line transcription is amongst the earliest studies on transcribing a single instrument track from a rich, polyphonic music signal [188], [189]. The problem of bass line transcription, which typically refers to the lower frequency range between 30 and 250 Hz, bears many similarities with the main melody extraction problem. Indeed, as for melody, the task is often regarded as a mere predominant-F0 estimation, where the tracking is now done in the lower frequency range. Not surprisingly, there are a number of approaches that were proposed to extract both melody and bass line within the same general framework [17], [174].

As an example, we now describe the system proposed in [190] in more detail. It is based on two frame-based feature extractors: namely a multiple-F0 estimator that provides salience value for four F0 estimates and an accent estimator that measures the probability of having an onset. One of the specificities of this approach is the integration of two different models: a note and a rest model. Bass notes are modeled using a three-state left-right HMM (the three states aim at capturing the attack, sustain and release phases of a played note) while rest notes are represented by a four-component GMM model (equivalent to a single state HMM). In subsequent work, a third background model was included to better represent the notes played by other instruments [174]. Another specificity of this model is the use of a musicological model that controls transition probabilities between the note models and the rest model. These transition probabilities depend on the musical key and on the preceding notes. The musical key is estimated from the set of four F0 candidates over the entire history of the piece in combination with simple artificial key profiles. The sequence model is either a bi-gram model or a more sophisticated variable-order Markov Model learned

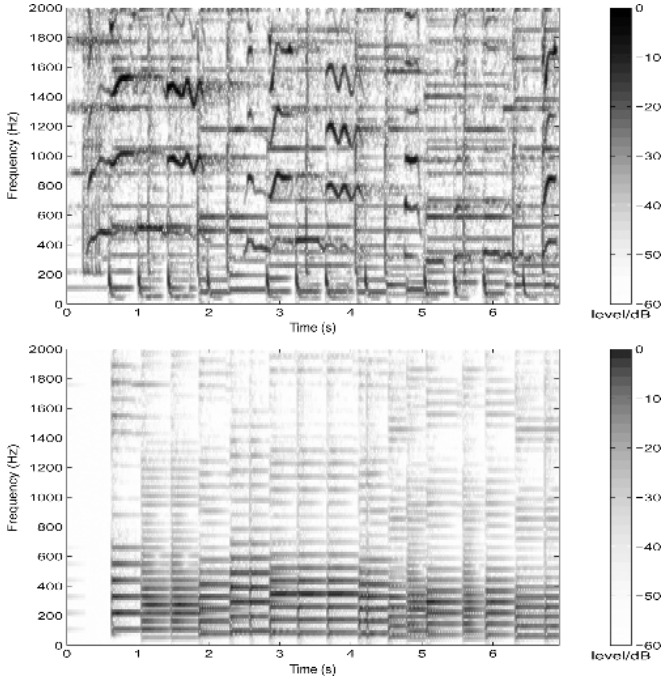

Fig. 16. Spectrograms of the original signal (top) and of the bass-line (bottom) resynthesized from its estimated MIDI transcription. From [190].

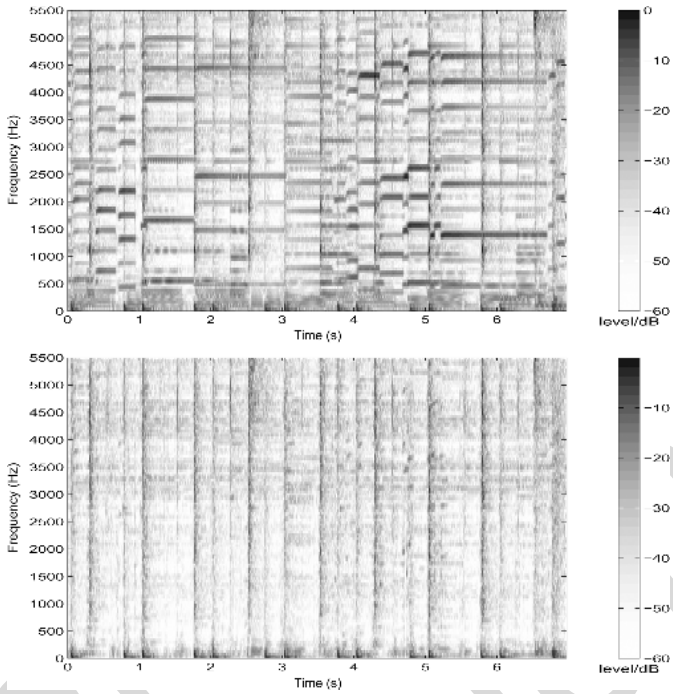

Fig. 17. Spectrograms of the original signal (top) and of the separated drum signal (bottom) obtained by enhanced Wiener filtering. From [144].

on a collection of MIDI files. Finally, Viterbi decoding is performed to obtain the most probable path through the models which yields a transcription of the bass line. As an illustration, Fig. 16 shows the spectrograms of a polyphonic music signal and of the bass-line resynthesized from its estimated MIDI transcription.

E. Drum Extraction

Historically, the analysis of the percussive or drum component of music signals has attracted less attention from the research community. Nevertheless, it has now been recognized that numerous applications can benefit from an appropriate processing of this percussive component including beat tracking, content-based retrieval based on the query-by-tapping paradigm, or beatboxing [191]–[194]. Furthermore, the rhythmic content, if successfully extracted, is crucial for developing music similarity measures as needed in music classification tasks. Even though some studies have focussed on traditional percussive instruments (see for example [195]–[197] for Indian percussions), most research has targeted the western drum kit composed of at least three main instrument classes (bass drum, snare drum, and cymbals). Starting from solo drum signals (see [198] for a review), most of the recent studies tackle the more realistic scenario of extracting and transcribing drum signals directly from polyphonic signals.

Following [144], the different approaches can be classified in three categories: *segment and classify*, *match and adapt*, or *separate and detect*. The *segment and classify* approaches either first segment the signal into individual discrete segments which are then classified using machine learning techniques [199]–[201] or jointly perform the two steps using, for example, hidden Markov models [202]. The *match and adapt* approaches rather aim at searching for occurrences of reference templates in the music signal which can be further adapted to the specificity of the analyzed signal [203], [204]. Finally, the *separate and detect* approaches first aim at separating or extracting the percussive component before further analysis to identify the drum events. As it has turned out, the approaches based on Independent Subspace Analysis (ISA) [205], [206], or [207] or non-negative matrix factorization [208], [209] are among the most successful systems for drum track separation.

To illustrate some of the above ideas, we now describe the drum track separation approach proposed in [144] in more detail. This algorithm is based on the general Wiener filtering approach for source separation, optimized for drum separation. The separation itself is performed using (13) but exploits different strategies for learning the spectral templates $\sigma_m^2(f)$ of each source. The drum spectral templates are learned on solo drum signals by non-negative matrix factorization (NMF) while the spectral templates for the background music are learned by a correlation-based clustering algorithm (as in [210]). A total of 144 templates are then learned—128 for the background music and 16 for the drum component. Then, an adaptation procedure is applied to better cope with the inherent variability of real audio signals. This adaptation consists in extending the set of 16 learned drum spectral templates by the PSD of the stochastic component obtained by subband subspace projection [211]. Indeed, this additional template already provides a decent estimate of the PSD of the drum signal and therefore facilitates the convergence of the algorithm to an appropriate solution. Finally, in order to represent at the same time the short drum onsets and steady part of tonal components, a multi-resolution approach is followed by implementing a window-size switching scheme for time–frequency decomposition based on the output of a note onset detection algorithm.

Fig. 17 gives an example of the result of the above algorithm by displaying the spectrograms of a polyphonic music signal and its separated drum signal.

VI. Conclusion

Signal processing for music analysis is a vibrant and rapidly evolving field of research, which can enrich the wider signal processing community with exciting applications and new problems. Music is arguably the most intricate and carefully constructed of any sound signal, and extracting information of relevance to listeners therefore requires the kinds of specialized methods that we have presented, able to take account of music-specific characteristics including pitches, harmony, rhythm, and instrumentation.

It is clear, however, that these techniques are not the end of the story for analyzing music audio, and many open questions and research areas remain to be more fully explored. Some of the most pressing and promising are listed below.

- Decomposing a complex music signal into different components can be a powerful preprocessing step for many applications. For example, since a cover song may preserve the original melody but alter the harmonization, or alternatively keep the same basic chord progression but devise a new melody (but rarely both), a promising approach to cover version recognition is to separate the main melody and accompaniment, and search in both parts independently [180]. A different decomposition, into harmonic and percussive components, has brought benefits to tasks such as chord recognition [22], genre classification [105], and beat tracking [61]. Note that such decompositions need not be perfect to yield benefits—even a modest improvement in the relative level between components can give a significant improvement in a subsequent analysis.
- Improved recognition and separation of sources in polyphonic audio remains a considerable challenge, with great potential to improve both music processing and much broader applications in computational auditory scene analysis and noise-robust speech recognition. In particular, the development, adaptation, and exploitation of sound source models for the purpose of source separation seems to be required in order to achieve an accuracy comparable to that of human listeners in dealing with polyphonic audio [212].
- In conjunction with the appropriate signal processing and representations, machine learning has had some great successes in music signal analysis. However, many areas are limited by the availability of high-quality labeled data. In chord recognition, for instance, the entire field is using the same corpus of 200 tracks for which high-quality manual chord transcripts have been prepared [23]. However, while special-purpose human labeling remains the gold standard, it is interesting to note that a given piece of music may have multiple, closely related sources of information, including alternate recordings or performances, partial mixes derived from the original studio multitracks, score representations including MIDI versions, lyric transcriptions, etc. These different kinds of information, some available in large quantities, present opportunities for innovative processing that can solve otherwise intractable problems such as score-guided separation [213], generate substitutes for manual ground-truth labels using music synchronization techniques [28], [29], [32], or use multi-perspective approaches to automatically evaluate algorithms [82], [214].
- Source separation and audio transcription, despite their obvious relationship, are often tackled as independent and separate tasks: As we have shown, a number of music signal analysis systems include some level of source separation. Other work in source separation has shown that performance is usually improved when appropriate prior information is used—information such as musical scores. Rather than relying on existing, ground-truth scores, information of this kind could also be obtained from rudimentary (or more elaborate) automatic transcription. Significant progress can perhaps be made in both fields by better exploiting transcription in source separation (so-called "informed" source separation) and by better integrating source separation in transcription systems.
- Many music analysis tasks have encountered a "glass ceiling," a point beyond which it has become very difficult to make improvements. One tactic is to restrict the domain, to allow an approach to specialize on a limited subset—for example, by building a beat tracker that is specialized for jazz, and a different one for classical music. This suggests a broader strategy of deploying a context-adaptation layer, able to choose parameters and models best suited to each particular signal. In the simplest case, this can be implemented by training the methods separately for, say, different genres, and then using automatic audio classification to choose the best models for a given test signal, but how to implement a more general and optimal context adaptation is a deep and open research question.
- Surprisingly, knowledge about auditory perception has a limited role is most music signal processing systems, but since music exists purely to be heard, hearing science promises to advance our understanding music perception and should therefore inform the analysis of complex signals such as polyphonic mixtures. In multipitch estimation, for example, understanding the way that overlapping partials interact and how their amplitudes can be precisely estimated represents one promising direction [176], [215].
- Much current research focuses on individual aspects of the music (e.g., the rhythm, or the chords, or the instruments). These aspects, however, are anything but independent, and we expect significant synergies from efforts to analyze them jointly, with information from one aspect helping to improve the extraction of another. Some examples of this include approaches that jointly use metric, harmonic, and structural cues to support and stabilizing tempo and beat tracking [20], [54], [75], [216], [217].

The work described in this paper illustrates the broad range of sophisticated techniques that have been developed in the rapidly evolving field of music signal analysis, but as shown by this list of open questions, there is much room for improvement and for new inventions and discoveries, leading to more powerful and innovative applications. While, for the moment, human listeners

remain far superior to machines in extracting and understanding the information in music signals, we hope that continued development of automatic techniques will lessen this gap, and may even help to clarify some aspects of how and why people listen to music.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful comments. Meinard Müller is supported by the Cluster of Excellence on Multimodal Computing and Interaction (MMCI) at Saarland University and the German Research Foundation (DFG) under grant MU 2686/3-1. Sebastian Ewert is supported by the German Research Foundation (DFG) under grant EW 104/4-1.

REFERENCES

[1] C. Roads, *The Computer Music Tutorial*. Cambridge, MA: MIT Press, 1996.

[2] M. S. Puckette, *The Theory and Technique of Electronic Music*. Singapore: World Scientific, 2007.

[3] J. A. Moorer, "On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer," Ph.D. dissertation, Department of Music, Stanford Univ., Stanford, CA, 1975.

[4] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. New York: Academic, 2003.

[5] E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models*. New York: Springer-Verlag, 1990.

[6] M. Müller, *Information Retrieval for Music and Motion*. New York: Springer-Verlag, 2007.

[7] K. L. Kashima and B. Mont-Reynaud, "The bounded-Q approach to time-varying spectral analysis," Dept. of Music, Stanford Univ., Tech. Rep. STAN-M-28, 1985.

[8] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant Q transform," *J. Acoust. Soc. Amer. (JASA)*, vol. 92, pp. 2698–2701, 1992.

[9] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. Sound Music Comput. Conf. (SMC)*, Barcelona, Pain, 2010.

[10] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proc. ICMC*, Beijing, China, 1999, pp. 464–467.

[11] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, 2001, pp. 15–18.

[12] D. P. W. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Honolulu, HI, Apr. 2007, vol. 4, pp. 1429–1439.

[13] M. Mauch and S. Dixon, "Approximate note transcription for the improved identification of difficult chords," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 135–140.

[14] M. Müller and S. Ewert, "Towards timbre-invariant audio features for harmony-based music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 649–662, Mar. 2010.

[15] A. Sheh and D. P. W. Ellis, "Chord segmentation and recognition using EM-trained hidden Markov models," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Baltimore, MD, 2003.

[16] E. Gómez, "Tonal description of music audio signals" Ph.D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2006 [Online]. Available: files/publications/emilia-PhD-2006.pdf

[17] 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. (ISCA J.)*, vol. 43, no. 4, pp. 311–329, 2004.

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

[19] K. Lee and M. Slaney, "Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 291–301, Feb. 2008.

[20] H. Papadopoulos and G. Peeters, "Simultaneous estimation of chord progression and downbeats from an audio file," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 121–124.

[21] M. Mauch, K. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Kobe, Japan, 2009, pp. 231–236.

[22] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama, "HMM-based approach for automatic chord detection using refined acoustic features," in *Proc. 35nd IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, 2010, pp. 5518–5521.

[23] C. Harte, M. Sandler, S. Abdallah, and E. Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005.

[24] S. Ewert, M. Müller, and R. B. Dannenberg, "Towards reliable partial music alignments using multiple synchronization strategies," in *Proc. Int. Workshop Adaptive Multimedia Retrieval (AMR)*, Madrid, Spain, Sep. 2009.

[25] D. Damm, C. Fremerey, F. Kurth, M. Müller, and M. Clausen, "Multimodal presentation and browsing of music," in *Proc. 10th Int. Conf. Multimodal Interfaces (ICMI)*, Chania, Crete, Greece, Oct. 2008, pp. 205–208.

[26] R. Dannenberg, "An on-line algorithm for real-time accompaniment," in *Proc. Int. Comput. Music Conf. (ICMC)*, 1984, pp. 193–198.

[27] M. Müller, H. Mattes, and F. Kurth, "An efficient multiscale approach to audio synchronization," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR)*, Victoria, BC, Canada, Oct. 2006, pp. 192–197.

[28] R. J. Turetsky and D. P. Ellis, "Ground-truth transcriptions of real music from force-aligned MIDI syntheses," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Baltimore, MD, 2003, pp. 135–141.

[29] N. Hu, R. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, Oct. 2003.

[30] S. Dixon and G. Widmer, "Match: A music alignment tool chest," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005.

[31] C. Fremerey, F. Kurth, M. Müller, and M. Clausen, "A demonstration of the SyncPlayer system," in *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR)*, Vienna, Austria, Sep. 2007, pp. 131–132.

[32] S. Ewert, M. Müller, and P. Grosche, "High resolution audio synchronization using chroma onset features," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 1869–1872.

[33] J. Serrà, E. Gómez, P. Herrera, and X. Serra, "Chroma binary similarity and local alignment applied to cover song identification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 6, pp. 1138–1151, Aug. 2008.

[34] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, pp. 195–197, 1981.

[35] M. Casey and M. Slaney, "Fast recognition of remixed music audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2007, pp. 1425–1428.

[36] M. Casey, C. Rhodes, and M. Slaney, "Analysis of minimum distances in high-dimensional musical spaces," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 1015–1028, Jul. 2008.

[37] F. Kurth and M. Müller, "Efficient index-based audio matching," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 382–395, Feb. 2008.

[38] D. B. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. Cambridge, MA: MIT Press, 2006.

[39] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *Proc. 11th Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 625–636.

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

[41] ***[AUTHOR: Please provide end page]***M. Müller and F. Kurth, "Towards structural analysis of audio recordings in the presence of musical variations," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, p. 163, 2007.

[42] M. Levy and M. Sandler, "Structural segmentation of musical audio by constrained clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 318–326, Feb. 2008.

[43] J. Paulus and A. Klapuri, "Music structure analysis using a probabilistic fitness measure and a greedy search algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1159–1170, Aug. 2009.

[44] R. J. Weiss and J. P. Bello, "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 123–128.

[45] T. Bertin-Mahieux, R. J. Weiss, and D. P. W. Ellis, "Clustering beat-chroma patterns in a large music database," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 111–116.

[46] R. Parncutt, "A perceptual model of pulse salience and metrical accent in musical rhythms," *Music Percept.*, vol. 11, pp. 409–464, 1994.

[47] W. A. Sethares, *Rhythm and Transforms*. New York: Springer, 2007.

[48] F. Lerdahl and R. Jackendoff, *Generative Theory of Tonal Music*. Cambridge, MA: MIT Press, 1983.

[49] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1035–1047, Sep. 2005.

[50] N. Collins, "A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions," in *Proc. AES Conv. 118*, Barcelona, Spain, 2005.

[51] R. Zhou, M. Mattavelli, and G. Zoia, "Music onset detection based on resonator time frequency image," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1685–1695, Nov. 2008.

[52] N. Collins, "Using a pitch detector for onset detection," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005, pp. 100–106.

[53] A. J. Eronen and A. P. Klapuri, "Music tempo estimation with k-NN regression," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 50–57, Jan. 2010.

[54] 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.

[55] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Sep. 2008.

[56] C. C. Toh, B. Zhang, and Y. Wang, "Multiple-feature fusion based onset detection for solo singing voice," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, 2009.

[57] A. P. Klapuri, A. J. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 342–355, Jan. 2006.

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

[59] P. Masri and A. Bateman, "Improved modeling of attack transients in music analysis-resynthesis," in *Proc. Int. Comput. Music Conf. (ICMC)*, Hong Kong, 1996, pp. 100–103.

[60] M. Alonso, B. David, and G. Richard, "Tempo and beat estimation of musical signals," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[61] M. Alonso, G. Richard, and B. David, "Accurate tempo estimation based on harmonic+noise decomposition," *EURASIP J. Adv. Signal Process.*, vol. 2007, p. 14, 2007, Article ID 82 795.

[62] P. Grosche and M. Müller, "A mid-level representation for capturing dominant tempo and pulse information in music recordings," in *Proc. 10th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, Oct. 2009, pp. 189–194.

[63] P. Grosche and M. Müller, "Extracting predominant local pulse information from music recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, 2011, to be published.

[64] 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. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1832–1844, Sep. 2006.

[65] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1009–1020, Mar. 2007.

[66] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, 2007.

[67] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, p. 158, 2007.

[68] J. Seppänen, A. Eronen, and J. Hiipakka, "Joint beat & tatum tracking from music signals," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Victoria, Canada, 2006.

[69] ***[AUTHOR: Please provide end page]***A. T. Cemgil, B. Kappen, P. Desain, and H. Honing, "On tempo tracking: Tempogram representation and Kalman filtering," *J. New Music Res.*, vol. 28, no. 4, pp. 259–273, 2001.

[70] K. Jensen, J. Xu, and M. Zachariasen, "Rhythm-based segmentation of popular chinese music," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005.

[71] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *Proc. Int. Conf. Multimedia Expo (ICME)*, Los Alamitos, CA, 2001.

[72] F. Kurth, T. Gehrmann, and M. Müller, "The cyclic beat spectrum: Tempo-related audio features for time-scale invariant audio identification," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR)*, Victoria, BC, Canada, Oct. 2006, pp. 35–40.

[73] P. Grosche, M. Müller, and F. Kurth, "Cyclic tempogram—A mid-level tempo representation for music signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, Mar. 2010, pp. 5522–5525.

[74] G. Peeters, "Time variable tempo detection and beat marking," in *Proc. Int. Comput. Music Conf. (ICMC)*, Barcelona, Spain, 2005.

[75] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *J. New Music Res.*, vol. 30, pp. 39–58, 2001.

[76] J. Seppänen, "Tatum grid analysis of musical signals," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2001, pp. 131–134.

[77] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Paris, France, 2002, pp. 150–156.

[78] N. Degara, A. Pena, M. E. P. Davies, and M. D. Plumbley, "Note onset detection using rhythmic structure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, 2010, pp. 5526–5529.

[79] J. Bilmes, "Techniques to foster drum machine expressivity," in *Proc. Int. Comput. Music Conf.*, Tokyo, Japan, 1993.

[80] F. Gouyon and P. Herrera, "Pulse-dependent analysis of percussive music," in *Proc. AES 22nd Int. Conf. Virtual, Synthetic Entertainment Audio*, Espoo, Finland, 2002.

[81] S. Dixon and W. Goebl, "Pinpointing the beat: Tapping to expressive performances," in *Proc. Int. Conf. Music Percep. Cogn.*, Sydney, Australia, 2002, pp. 617–620.

[82] P. Grosche, M. Müller, and C. S. Sapp, "What makes beat tracking difficult? A case study on chopin mazurkas," in *Proc. 11th Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 649–654.

[83] E. W. Large and C. Palmer, "Perceiving temporal regularity in music," *Cognitive Sci.*, vol. 26, no. 1, pp. 1–37, 2002.

[84] S. Dixon, F. Gouyon, and G. Widmer, "Towards characterization of music via rhythmic patterns," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[85] G. Peeters, "Rhythm classification using spectral rhythm patterns," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005, pp. 644–647.

[86] F. Gouyon and P. Herrera, "Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *AES Conv. 114*, Amsterdam, The Netherlands, 2003.

[87] S. Dixon, "Evaluation of the audio beat tracking system beatroot," *J. New Music Res.*, vol. 36, pp. 39–50, 2007.

[88] USA Standard Acoustical Terminology American National Standards Inst., Tech. Rep. S1.1-1960, 1960.

[89] J.-J. Aucouturier, "Dix expériences sur la modélisation du timbre polyphonique [Ten Experiments on the Modelling of Polyphonic Timbre]," Ph.D. dissertation, Univ. Paris 6, Paris, France, 2006.

[90] V. Alluri and P. Toiviainen, "Exploring perceptual and acoustical correlates of polyphonic timbre," *Music Percept.*, vol. 27, no. 3, pp. 223–241, 2010.

[91] R. Erickson, *Sound Structure in Music*. Berkeley, CA: Univ. of California, 1975.

[92] J. Barbour, "Analytic listening: A case study of radio production," in *Proc. Int. Conf. Auditory Display*, Sydney, Australia, Jul. 2004.

[93] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 133–141, Mar. 2006.

[94] Y. E. Kim, E. M. Schmidt, R. Migneco, B. C. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. 11th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 255–266.

[95] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 467–476, Mar. 2008.

[96] J. F. Schouten, "The perception of timbre," in *Proc. 6th Int. Congr. Acoust.*, Tokyo, Japan, 1968, p. GP-6-2.

[97] S. Handel, "Timbre perception and auditory object identification," in *Hearing—Handbook of Perception and Cognition*, B. C. J. Moore, Ed., 2nd ed. San Diego, CA: Academic, 1995, pp. 425–460.

[98] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," *J. Acoust. Soc. Amer.*, vol. 118, no. 1, pp. 471–482, 2005.

[99] R. Cogan, *New Images of Musical Sound*. Cambridge, MA: Harvard Univ. Press, 1984.

[100] R. A. Kendall and E. C. Carterette, "Perceptual scaling of simultaneous wind instrument timbres," *Music Percept.*, vol. 8, no. 4, pp. 369–404, 1991.
[101] A. Eronen, "Comparison of features for musical instrument recognition," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, 2001, pp. 19–22.
[102] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, ser. Signal Processing Series. Englewood Cliffs: Prentice-Hall, 1993.
[103] X. Serra, "Musical sound modeling with sinusoids plus noise," in *Musical Signal Processing*, C. Roads, S. Pope, A. Picialli, and G. D. Poli, Eds. Lisse, The Netherlands: Swets & Zeitlinger, 1997.
[104] N. Ono, K. Miyamoto, J. LeRoux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. Eur. Signal Process. Conf.*, Lausanne, Switzerland, 2008, pp. 240–244.
[105] H. Rump, S. Miyabe, E. Tsunoo, N. Ono, and S. Sagama, "Autoregressive MFCC models for genre classification improved by harmonic-percussion separation," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 87–92.
[106] *Hearing—Handbook of Perception and Cognition*, B. C. J. Moore, Ed., 2nd ed. San Diego, CA: Academic, 1995.
[107] M. S. Vinton and L. E. Atlas, "Scalable and progressive audio codec," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, 2001, pp. 3277–3280.
[108] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 576–588, Mar. 2010.
[109] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Munich, Germany, 1997, pp. 209–212.
[110] A. Cont, S. Dubnov, and D. Wessel, "Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negativity constraints," in *Proc. 10th Int. Conf. Digital Audio Effects*, Bordeaux, France, 2007, pp. 85–92.
[111] V. Välimäki, J. Pakarinen, C. Erkut, and M. Karjalainen, "Discrete-time modelling of musical instruments," *Rep. Progr. Phys.*, vol. 69, no. 1, pp. 1–78, 2006.
[112] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Tampa, FL, 1985, pp. 937–940.
[113] A. Klapuri, "Analysis of musical instrument sounds by source-filter-decay model," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, Honolulu, HI, 2007, pp. 53–56.
[114] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 564–575, Mar. 2010.
[115] T. Heittola, A. Klapuri, and T. Virtanen, "Musical instrument recognition in polyphonic audio using source-filter model for sound separation," in *Proc. 10th Int. Soc. Music Inf. Retrieval Conf. (ISMIR)*, Kobe, Japan, 2009, pp. 327–332.
[116] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale*. London, U.K.: Springer, 1998.
[117] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.
[118] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[119] P. Herrera-Boyer, A. Klapuri, and M. Davy, "Automatic classification of pitched musical instrument sounds," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2006, pp. 163–200.
[120] T. Kitahara, "Computational musical instrument recognition and its application to content-based music information retrieval," Ph.D. dissertation, Kyoto Univ., , Japan, Mar. 2007.
[121] A. Eronen, "Signal processing method for audio classification and music content analysis," Ph.D. dissertation, Tampere Univ. of Technol., Tampere, Finland, Jun. 2009.
[122] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 1, pp. 174–186, 2009.
[123] A. A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Proc. Int. Conf. Digital Audio Effects*, Naples, Italy, 2004.
[124] S. Essid, G. Richard, and B. David, "Instrument recognition in polyphonic music based on automatic taxonomies," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 68–80, 2006.
[125] G. Peeters, "A Large set of audio features for sound description (similarity and classification) in the CUIDADO Project," IRCAM, Paris, France, Apr. 2004, Tech. Rep..
[126] S. Essid, "Classification automatique des signaux audio-fréquences: Reconnaissance des instruments de musique," Ph.D. dissertation, Univ. Pierre et Marie Curie, Paris, France, Dec. 2005.
[127] A. Eronen, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs," in *Proc. th Int. Symp. Signal Process. and Its Applicat.*, Paris, France, 2003, pp. 133–136.
[128] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Musical instrument recognizer "instrogram" and its application to music retrieval based on instrument similarity," in *Proc. IEEE Int. Symp. Multimedia*, San Diego, CA, 2006, pp. 265–272.
[129] D. Little and B. Pardo, "Learning musical instruments from mixtures of audio with weak labels," in *Proc. 9th Int. Symp. Music Inf. Retrieval (ISMIR)*, Philadelphia, PA, 2008.
[130] J. Eggink and G. J. Brown, "A missing feature approach to instrument identification in polyphonic music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Hong Kong, China, 2003, pp. 553–556.
[131] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.
[132] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.
[133] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps," *EURASIP J. Appl. Signal Process.*, vol. 2007, pp. 1–15, 2007.
[134] B. Kostek, "Musical instrument recognition and duet analysis employing music information retrieval techniques," *Proc. IEEE*, vol. 92, pp. 712–729, 2004.
[135] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic instrument recognition using spectral clustering," in *Proc. 8th Int. Symp. Music Inf. Retrieval*, Vienna, Austria, 2007.
[136] J. J. Burred, A. Röbel, and T. Sikora, "Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, 2009, pp. 173–176.
[137] , D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York: Wiley/IEEE Press, 2006.
[138] E. Vincent and X. Rodet, "Instrument identification in solo and ensemble music using independent subspace analysis," in *Proc. 5th Int. Symp. Music Inf. Retrieval*, Barcelona, Spain, 2004, pp. 576–581.
[139] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 116–128, Jan. 2008.
[140] K. Kashino and H. Murase, "A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Commun.*, vol. 27, pp. 337–349, 1999.
[141] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals," in *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2006, ch. 6, pp. 267–296.
[142] P. Comon and C. Jutten, *Handbook of Blind Source Separation, Independent Component Analysis and Applications*. New York: Academic, Elsevier, 2010.
[143] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Stanford Univ. , Stanford, CA, 1989.
[144] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 529–540, Mar. 2008.
[145] F. B. Laurent Benaroya and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 191–199, Jan. 2006.
[146] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[147] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[148] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[149] T. Virtanen and A. Klapuri, "Analysis of polyphonic audio using source-filter model and non-negative matrix factorization,". Whistler, Canada, 2006, extended abstract.

[150] R. G. Emmanuel Vincent and C. Féevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2005.

[151] W. Hess, *Pitch Determination of Speech Signals*. Berlin, Germany: Springer-Verlag, 1983.

[152] W. Hess, *Pitch and Voicing Determination*. New York: Marcel Dekker, 1992, pp. 3–48.

[153] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *J. Acoust. Soc. Amer. (JASA)*, vol. 43, no. 4, pp. 829–834, 1968.

[154] P. Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 1982, pp. 180–183.

[155] J.-S. Liénard, F. Signol, and C. Barras, "Speech fundamental frequency estimation using the alternate comb," in *Proc. Interspeech*, Antwerpen, Belgium, 2007.

[156] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer. (JASA)*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[157] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, ch. 14, pp. 495–518.

[158] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[159] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *J. Acoust. Soc. Amer. (JASA)*, vol. 89, no. 6, pp. 2866–2882, 1991.

[160] G. Peeters, "Music pitch representation by periodicity measures based on combined temporal and spectral representations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, pp. 53–56.

[161] A. de Cheveigne, "Multiple f0 estimation," in *Computational Auditory Scene Analysis*, D. Wang and G. J. Brown, Eds. New York: Wiley/ IEEE Press, 2006.

[162] *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2006.

[163] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," in *Synthesis Lectures on Speech Audio Process.*. San Rafael, CA: Morgan and Claypool, 2009.

[164] R. Badeau, V. Emiya, and B. David, "Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra," in *Proc. ICASSP'09*, Taipei, Taiwan, Apr. 2009, pp. 3073–3076.

[165] C. Yeh, "Multiple fundamental frequency estimation of polyphonic recordings," Ph.D. dissertation, Univ. Pierre et Marie Curie (Paris 6), Paris, France, 2008.

[166] F. Signol, C. Barras, and J.-S. Lienard, "Evaluation of the pitch estimation algorithms in the monopitch and multipitch cases," in *Proc. Acoustics'08*, 2008, vol. 123, no. 5, p. 3077.

[167] ***[AUTHOR: Please provide page range]***G. E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, 2007.

[168] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 439–449, Jun. 2004.

[169] J. P. Bello, L. Daudet, and M. Sandler, "Automatic piano transcription using frequency and time-domain information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2242–2251, Aug. 2006.

[170] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2003, pp. 177–180.

[171] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

[172] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[173] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[174] M. Ryynänen and A. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Comput. Music J.*, vol. 32, no. 3, pp. 72–86, 2008.

[175] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 255–266, Feb. 2008.

[176] C. Yeh and A. Roebel, "The expected amplitude of overlapping partials of harmonic sounds," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP'09)*, Taipei, Taiwan, 2009, pp. 316–319.

[177] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: Azimuth discriminiation and resynthesis," in *Proc. 7th Int. Conf. Digital Audio Effects (DAFX-04)*, Naples, Italy, Oct. 2004.

[178] S. Pauws, "CubyHum: A fully operational query by humming system," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Paris, France, 2002.

[179] J. Weil, J.-L. Durrieu, G. Richard, and T. Sikora, "Automatic generation of lead sheets from polyphonic music signals," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Kobe, Japan, Oct. 2009, pp. 603–608.

[180] R. Foucard, J.-L. Durrieu, M. Lagrange, and G. Richard, "Multimodal similarity between musical streams for cover version detection," in *Proc. ICASSP*, Dallas, TX, Mar. 2010, pp. 5514–5517.

[181] R. P. Paiva, "Melody detection in polyphonic audio," Ph.D. dissertation, Univ. of Coimbra, Coimbra, Portugal, 2006.

[182] M. Goto, "A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2000, vol. 2, pp. 757–760.

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

[184] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," in *Proc. 9th Int. Conf. Digital Audio Effects (DAfX-06)*, 2006, pp. 18–20.

[185] K. Dressler, "Audio melody extraction," in *Proc. Int. Soc. Music Inf. Retrieval Conf. (ISMIR): Late Breaking Session*, 2010.

[186] J.-L. Durrieu, "Transcription et séparation automatique de la mélodie principale dans les signaux de musique polyphoniques," Ph.D. dissertation, Télécom ParisTech, Paris, France, May 2010.

[187] E. Tsunoo, T. Akase, N. Ono, and S. Sagayama, "Music mood classification by rhythm and bass-line unit pattern analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2010, pp. 265–268.

[188] M. Goto and S. Hayamizu, "A real-time music scene description system: Detecting melody and bass lines in audio signals," in *Proc. Int. Workshop Computat. Auditory Scene Anal.*, Aug. 1999.

[189] S. Hainsworth and M. D. Macleod, "Automatic bass line transcription from polyphonic music," in *Proc. Int. Comput. Music Conf.*, Havana, Cuba, 2001.

[190] M. Ryynänen and A. Klapuri, "Automatic bass line transcription from streaming polyphonic audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Honolulu, HI, 2007, pp. 1437–1440.

[191] A. Kapur, M. Benning, and G. Tzanetakis, "Query-by-beat-boxing: Music retrieval for the dj," in *Proc. 6th Int. Conf. Music Inf. Retrieval, (ISMIR)*, Barcelona, Spain, Sep. 2004.

[192] T. Nakano, J. Ogata, M. Goto, and Y. Hiraga, "A drum pattern retrieval method by voice percussion," in *Proc. 6th Int. Conf. Music Inf. Retrieval, (ISMIR)*, 2004, pp. 550–553.

[193] J. C. C. Chen and A. L. Chen, "Query by rhythm: An approach for song retrieval in music databases," in *Proc. Int. Workshop Res. Iss. Data Eng.*, Feb. 1998.

[194] O. Gillet and G. Richard, "Drum loops retrieval from spoken queries," *J. Intell. Inf. Syst.—Special Iss. Intell. Multimedia Applicat.*, vol. 24, no. 2/3, pp. 159–177, Mar. 2005.

[195] P. Chordia, "Automatic transcription of solo tabla music," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2006.

[196] P. Chordia and A. Rae, "Tabla gyan: A system for realtime tabla recognition and resynthesis," in *Proc. Int. Comput. Music Conf. (ICMC)*, 2008.

[197] O. Gillet and G. Richard, "Automatic e.g., ing of tabla signals," in *Proc. 4th ISMIR Conf. 2003*, Baltimore, MD, Oct. 2003.

[198] D. FitzGerald and J. Paulus, "Unpitched percussion transcription," in *Signal Processing Methods for Music Transcription*. New York: Springer, 2006, pp. 131–162.
[199] V. Sandvold, F. Gouyon, and P. Herrera, "Percussion classification in polyphonic audio recordings using localized sound models," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, Oct. 2004, pp. 537–540.
[200] K. Tanghe, S. Degroeve, and B. D. Baets, "An algorithm for detecting and labeling drum events in polyphonic music," in *Proc. 1st MIREX*, London, U.K., Sep. 2005.
[201] O. Gillet and G. Richard, "Drum track transcription of polyphonic music signals using noise subspace projection," in *Proc. 6th Int. Conf. Music Inf. Retrieval, (ISMIR)*, London, U.K., Sep. 2005.
[202] J. Paulus, "Acoustic modelling of drum sounds with hidden markov models for music transcription," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2006, pp. 241–244.
[203] A. Zils, F. Pachet, O. Delerue, and F. Gouyon, "Automatic extraction of drum tracks from polyphonic music signals," in *Proc. Int. Conf. Web Delivering of Music (WEDELMUSIC)*, Dec. 2002.
[204] K. Yoshii, M. Goto, and H. G. Okuno, "Automatic drum sound description for real-world music using template adaptation and matching methods," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2004.
[205] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *Proc. Int. Symp. Ind. Compon. Anal. Blind Signal Separat. (ICA)*, Apr. 2003.
[206] C. Uhle and C. Dittmar, "Further steps towards drum transcription of polyphonic music," in *Proc. 11th AES Conv.* , May 2004.
[207] D. FitzGerald, E. Coyle, and B. Lawlor, "Sub-band independent subspace analysis for drum transcription," in *Proc. 5th Int. Conf. Digital Audio Effects (DAFX-02)*, 2002.
[208] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using nonnegative matrix factorization and support vector machine," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2005.
[209] J. Paulus and T. Virtanen, "Drum transcription with non-negative spectrogram factorisation," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, 2005.
[210] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval, "Non negative sparse representation for wiener based source separation with a single sensor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2003, pp. 613–616.
[211] R. Badeau, B. David, and R. Boyer, "Eds parametric modeling and tracking of audio signals," in *Proc. 5th Int. Conf. Digital Audio Effects (DAFX-02)*, 2002, pp. 26–28.
[212] D. P. W. Ellis, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds. New York: Wiley/IEEE Press, 2006, ch. 4, pp. 115–146.
[213] J. Woodruff, B. Pardo, and R. B. Dannenberg, "Remixing stereo music with score-informed source separation," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, 2006, pp. 314–319.
[214] V. Konz, M. Müller, and S. Ewert, "A multi-perspective evaluation framework for chord recognition," in *Proc. 11th Int. Conf. Music Inf. Retrieval (ISMIR)*, Utrecht, The Netherlands, 2010, pp. 9–14.
[215] M. R. Every and J. E. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1845–1856, Sep. 2006.
[216] R. B. Dannenberg, "Toward automated holistic beat tracking, music analysis and understanding," in *Proc. Int. Conf. Music Inf. Retrieval (ISMIR)*, London, U.K., 2005, pp. 366–373.
[217] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1280–1289, Aug. 2010.

Meinard Müller (M'09) received the Diplom degree in mathematics, the Ph.D. degree in computer science, and the Habilitation degree in the field of multimedia retrieval from Bonn University, Bonn, Germany.

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

Daniel P. W. Ellis (M'96–SM'04) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He is an Associate Professor in the Electrical Engineering Department, Columbia University, New York. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with all aspects of extracting high-level information from audio, including speech recognition, music description, and environmental sound processing. He also runs the AUDITORY e-mail list of 1700 worldwide researchers in perception and cognition of sound. He worked at MIT, where he was a Research Assistant at the Media Lab, and he spent several years as a Research Scientist at the International Computer Science Institute, Berkeley, CA, where he remains an external fellow.

Anssi Klapuri (M'06) received the Ph.D. degree from Tampere University of Technology (TUT), Tampere, Finland, in 2004.

He visited as a Post-Doctoral Researcher at the Ecole Centrale de Lille, Lille, France, and Cambridge University, Cambridge, U.K., in 2005 and 2006, respectively. He worked until 2009 as a Professor at TUT. In December 2009, he joined Queen Mary, University of London, London, U.K., as a Lecturer in sound and music processing. His research interests include audio signal processing, auditory modeling, and machine learning.

Gaël Richard (M'02–SM'06) received the State Engineering degree from TELECOM ParisTech (formerly ENST), Paris, France, in 1990, the Ph.D. degree in speech synthesis from LIMSI-CNRS, University of Paris-XI, in 1994 and the *Habilitation à Diriger des Recherches* degree from the University of Paris XI in September 2001.

After the Ph.D. degree, he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches to speech production. Between 1997 and 2001, he successively worked for Matra Nortel Communications, Bois d'Arcy, France, and for Philips Consumer Communications, Montrouge, France. In particular, he was the Project Manager of several large-scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined the Department of Signal and Image Processing, TELECOM ParisTech, where he is now a Full Professor in audio signal processing and Head of the Audio, Acoustics, and Waves Research Group. He is coauthor of over 80 papers, inventor of a number of patents, and one of the experts of the European commission in the field of speech and audio signal processing.

Prof. Richard is a member of EURASIP and an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.