# Introduction to Music Transcription

Anssi Klapuri

Institute of Signal Processing, Tampere University of Technology,
Korkeakoulunkatu 1, 33720 Tampere, Finland
Anssi.Klapuri@tut.fi

Music transcription refers to the analysis of an acoustic musical signal so as to write down the pitch, onset time, duration, and source of each sound that occurs in it. In Western tradition, written music uses *note symbols* to indicate these parameters in a piece of music. Figures 1–2 show the notation of an example music signal. Omitting the details, the main conventions are that time flows from left to right and the pitch of the notes is indicated by their vertical position on the staff lines. In the case of drums and percussions, the vertical position indicates the instrument and the stroke type. The loudness (and the applied instrument in the case of pitched instruments) is normally not specified for individual notes but is determined for larger parts.

Besides the common musical notation, the transcription can take many other forms, too. For example, a guitar player may find it convenient to read *chord symbols* which characterize the note combinations to be played in a more general manner. In a computational transcription system, a MIDI file[1] is often an appropriate format for musical notations (Fig. 3). Common to all these representations is that they capture musically meaningful parameters that can be used in performing or synthesizing the piece of music in question. From this point of view, music transcription can be seen as discovering the "recipe", or, reverse-engineering the "source code" of a music signal.

A complete transcription would require that the pitch, timing, and instrument of all the sound events is resolved. As this can be very hard or even theoretically impossible in some cases, the goal is usually redefined as being either to notate as many of the constituent sounds as possible (complete transcription) or to transcribe only some well-defined part of the music signal, for example the dominant melody or the most prominent drum sounds (partial transcription). Both of these goals are relevant and are discussed in this book.

Music transcription is closely related to *structured audio coding*. A musical notation or a MIDI-file is an extremely compact representation yet retains the

[1]Musical Instrument Digital Interface (MIDI) is a standard for exchanging performance data and parameters between electronic musical devices [2, 3].

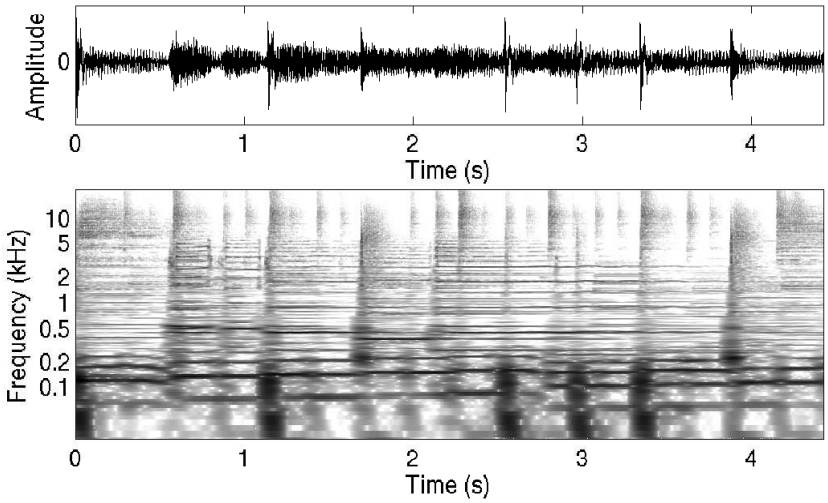


**Fig. 1.** An acoustic musical signal (top) and its time-frequency domain representation (bottom). The excerpt is from Song G034 in the RWC database [1].

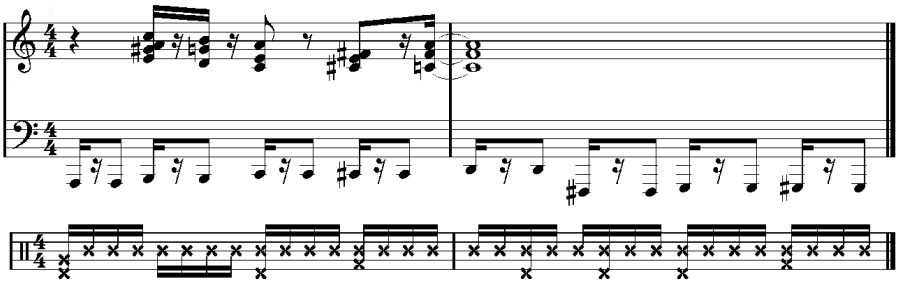

**Fig. 2.** Musical notation corresponding to the signal in Fig. 1. The upper staff lines show the notation for pitched musical instruments and the lower staff lines show the notation for percussion instruments.

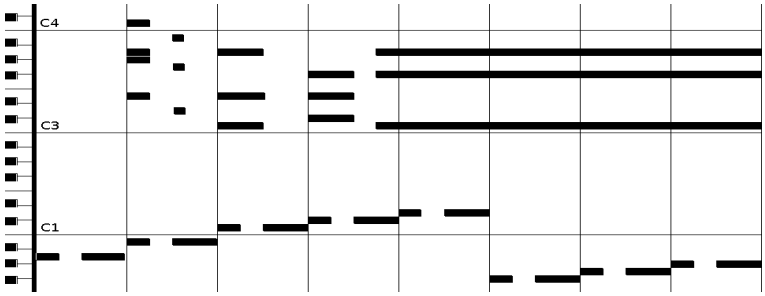


**Fig. 3.** A "piano-roll" illustration of a MIDI-file which corresponds to the pitched instruments in the signal in Fig. 1. Different notes are arranged on the vertical axis and time flows from left to right.

characteristics of a piece of music to an important degree. Another related area of study is that of *music perception* [4]. Detecting and recognizing individual sounds in music is a big part of its perception, although it should be emphasized that musical notation is primarily designed to serve sound production and not to model hearing. We do not hear music in terms of note symbols but, as described by Bregman [5, pp. 457–460], music often "fools" the auditory system so that we perceive simultaneous sounds as a single entity.

In addition to audio coding, applications of music transcription comprise

- *Music information retrieval* based on the melody of a piece, for example.
- *Music processing*, such as changing the instrumentation, arrangement, or the loudness of different parts before resynthesizing a piece from its score.
- *Human-computer interaction* in various applications, including score typesetting programs and musically-oriented computer games. Singing transcription is of particular importance here.
- *Music-related equipment*, ranging from music-synchronous light effects to highly sophisticated interactive music systems which generate an accompaniment for a soloist.
- *Musicological analysis* of improvised and ethnic music for which musical notations do not exist.
- *Transcription tools* for amateur musicians who wish to play along with their favourite music.

The purpose of this book is to describe algorithms and models for the different subtopics of music transcription, including pitch analysis, meter analysis (see Sect. 1 for term definitions), percussion transcription, musical instrument classification, and music structure analysis. The main emphasis is laid on the low-level signal analysis where sound events are detected and their parameters are estimated, and not so much on the subsequent processing of the note data to obtain larger musical structures. The theoretical background of different signal analysis methods is presented and their application to the transcription problem is discussed.

The primary target material considered in this book is complex music signals where several sounds are played simultaneously. These are referred to as *polyphonic* signals, in contrast to *monophonic* signals where at most one note is sounding at a time. For practical reasons, the scope is limited to Western music, although not to any particular genre. Many of the analysis methods make no assumptions about the larger-scale structure of the signal and are thus applicable to the analysis of music from other cultures as well.

To give a reasonable estimate of the achievable goals in automatic music transcription, it is instructive to study what human listeners are able to do in this task. An average listener perceives a lot of musically relevant information in complex audio signals. He or she can tap along with the rhythm, hum the melody (more or less correctly), recognize musical instruments, and locate structural parts of the piece, such as the chorus and the verse in popular music. Harmonic changes and various details are perceived less consciously.

Similarly to natural language, however, reading and writing music requires education. Not only the used notation needs to be studied, but recognizing different pitch intervals and timing relationships is an ability that has to be learned – these have to be encoded into a symbolic form in one's mind before writing them down. Moreover, an untrained listener is typically not able to hear out the inner lines in music (sub-melodies other than the dominant one) but musical ear training is needed to develop an analytic mode of listening where these can be distinguished. The richer the polyphonic complexity of a musical composition, the more its transcription requires musical ear training and knowledge of the particular musical style and of the playing techniques of the instruments involved.

First attempts towards the automatic transcription of polyphonic music were made in the 1970s, when Moorer proposed a system for transcribing two-voice compositions [6, 7]. His work was followed by that of Chafe et al. [8], Piszczalski [9], and Maher [10, 11] in the 1980s. In all these early systems, the number of concurrent voices was limited to two and the pitch relationships of simultaneous sounds were restricted in various ways. On the rhythm analysis side, the first algorithm for beat-tracking[2] in general audio signals was proposed by Goto and Muraoka in the 1990s [12], although this was preceded by a considerable amount of work for tracking the beat in parametric note data (see [13] for a summary) and by the beat tracking algorithm of Schloss for percussive audio tracks [14]. First attempts to transcribe percussive instruments were made in the mid-1980s by Schloss [14] and later by Bilmes [15], both of whom classified different types of conga strikes in continuous recordings. Transcription of polyphonic percussion tracks was later addressed by Goto and Muraoka [16]. A more extensive description of the early stages of music transcription has been given by Tanguiane in [17, pp. 3–6].

Since the beginning of 1990s, the interest in music transcription has grown rapidly and it is not possible to make a complete account of the work here. However, certain general trends and successful approaches can be discerned. One of these has been the use of *statistical methods*. To mention a few examples, Kashino [18], Goto [19], Davy and Godsill [20], and Ryynänen [21] proposed statistical methods for the pitch analysis of polyphonic music; in beat tracking, statistical methods were employed by Cemgil and Kappen [22], Hainsworth and MacLeod [23], and Klapuri et al. [24]; and in percussive instrument transcription by Gillet and Richard [25] and Paulus et al. [26]. In musical instrument classification, statistical pattern recognition methods prevail [27]. Another trend has been the increasing utilization of *computational models of the human auditory system*. These were first used for music transcription by Martin [28], and auditorily-motivated methods have since then been proposed for polyphonic pitch analysis by Karjalainen and Tolonen [29] and Klapuri [30], and for beat tracking by Scheirer [31], for example. Another

[2] *Beat tracking* refers to the estimation of a rhythmic pulse which corresponds to the tempo of a piece and (loosely) to the foot-tapping rate of human listeners.

prominent approach has been to model the human *auditory scene analysis* (ASA) ability. The term ASA refers to the way in which humans organize spectral components to their respective sounds sources and recognize simultaneously occurring sounds [5]. The principles of ASA were brought to the pitch analysis of polyphonic music signals by Mellinger [32] and Kashino [33], and later by Godsmark and Brown [34] and Sterian [35]. Most recently, several *unsupervised learning* methods have been proposed where a minimal amount of prior assumptions is made about the analyzed signal. Methods based on independent component analysis [36] were introduced to music transcription by Casey [37, 38], and various other methods have been later proposed by Lepain [39], Smaragdis [40, 41], Abdallah [42, 43], Virtanen (see Chapt. **??**), FitzGerald [44, 45], and Paulus [46]. Of course, there are also methods that do not represent any of the above-mentioned trends, and a more comprehensive review of the literature is presented in the coming chapters.

The state-of-the-art music transcription systems are still clearly inferior to skilled human musicians in accuracy and flexibility. That is, a reliable general-purpose transcription system does not exist at the present time. However, some degree of success has been achieved for polyphonic music of limited complexity. In the transcription of pitched instruments, typical restrictions are that the number of concurrent sounds is limited [29, 20], interference of drums and percussive sounds is not allowed [47], or only a specific instrument is considered [48]. Some promising results for the transcription of real-world music on CD recordings has been demonstrated by Goto [19] and Ryynänen and Klapuri [21]. In percussion transcription, quite good accuracy has been achieved in the transcription of percussive tracks which comprise a limited number of instruments (typically bass drum, snare, and hihat) and no pitched instruments [25, 46]. Also promising results have been reported for the transcription of the bass and snare drums on real-world recordings, but this is a more open problem (see e.g. Zils et al. [49], FitzGerald et al. [50], Yoshii et al. [51]). Beat tracking of complex real-world audio signals can be performed quite reliably with the state-of-the-art methods, but difficulties remain especially in the analysis of classical music and rhythmically complex material. Comparative evaluations of beat-tracking systems can be found in [23, 24, 52]. Research on musical instrument classification has mostly concentrated on working with isolated sounds, although more recently this has been attempted in polyphonic audio signals, too [53, 54, 55, 56].

## 1 Terminology and Concepts

Before turning to a more general discussion of the music transcription problem and the contents of this book, it is necessary to introduce some basic terminology of auditory perception and music. To discuss music signals, we first have to discuss the perceptual attributes of sounds of which they consist. There are

four subjective qualities that are particularly useful in characterizing sound events: pitch, loudness, duration, and timbre [57].

*Pitch* is a perceptual attribute which allows the ordering of sounds on a frequency-related scale extending from low to high. More exactly, pitch is defined as the frequency of a sine wave that is matched to the target sound by human listeners [58]. *Fundamental frequency* (F0) is the corresponding physical term and is defined for periodic or nearly periodic sounds only. For these classes of sounds, F0 is defined as the inverse of the period and is closely related to pitch. In ambiguous situations, the period corresponding to the perceived pitch is chosen.

The perceived *loudness* of an acoustic signal has a non-trivial connection to its physical properties, and computational models of loudness perception constitute a fundamental part of psychoacoustics[3] [59]. In music processing, however, it is often more convenient to express the level of sounds with their mean-square power and to apply a logarithmic (deciBel) scale to deal with the wide dynamic range involved. The perceived *duration* of a sound has more or less one-to-one mapping to its physical duration in cases where this can be unambiguously determined.

*Timbre* is sometimes referred to as sound "colour" and is closely related to the recognition of sound sources [61]. For example, the sounds of the violin and the flute may be identical in their pitch, loudness, and duration, but are still easily distinguished by their timbre. The concept is not explained by any simple acoustic property but depends mainly on the coarse spectral energy distribution of a sound, and the time evolution of this. Whereas pitch, loudness, and duration can be quite naturally encoded into a single scalar value, timbre is essentially a multidimensional concept and is typically represented with a feature *vector* in musical signal analysis tasks.

Musical information is generally encoded into the *relationships* between individual sound events and between larger entities composed of these. Pitch relationships are utilized to make up melodies and chords. Timbre and loudness relationships are used to create musical form especially in percussive music, where pitched musical instruments are not necessarily employed at all. Inter-onset interval (IOI) relationships, in turn, largely define the rhythmic characteristics of a melody or a percussive sound sequence (the term IOI refers to the time interval between the beginnings of two sound events). Although durations of the sounds play a role too, the IOIs are more crucial in determining the perceived rhythm [62]. Indeed, many rhythmically important instruments, such as drums and percussions, produce exponentially-decaying waveshapes that do not even have a uniquely defined duration. In the case of sustained musical sounds, however, the durations are used to control *articu-*

[3]Psychoacoustics is the science that deals with the perception of sound. In a psychoacoustic experiment, the relationships between an acoustic stimulus and the resulting subjective sensation is studied by presenting specific tasks or questions to human listeners [57].

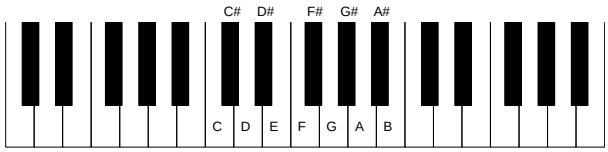


**Fig. 4.** Illustration of the piano keyboard (only three octaves are shown here).

*lation*. The two extremes here are "staccato", where notes are cut very short, and "legato", where no perceptible gaps are left between successive notes.

A *melody* is a series of pitched sounds with musically meaningful pitch and IOI relationships. In written music, this corresponds to a sequence of single notes. A *chord* is a combination of two or more simultaneous notes. A chord can be harmonious or dissonant, subjective attributes related to the specific relationships between the component pitches and their overtone partials. *Harmony* refers to the part of music theory which studies the formation and relationships of chords.

Western music arranges notes on a quantized logarithmic scale, with 12 notes in each octave range. The nominal fundamental frequency of note $n$ can be calculated as $440\ \text{Hz} \times 2^{n/12}$, where 440 Hz is an agreed-upon anchor point for the tuning and $n$ varies from $-48$ to 39 on a standard piano keyboard, for example. According to a musical convention, the notes in each octave are lettered as C, C#, D, D#, E, F, ... (see Fig. 4) and the octave is indicated with a number following this, for example A4 and A3 referring to the notes with fundamental frequencies 440 Hz and 220 Hz, respectively.

There are of course instruments which produce arbitrary pitch values and not just discrete notes like the piano. When playing the violin or singing, for example, both intentional and unintentional deviations take place from the nominal note pitches. In order to write down the music in a symbolic form, it is necessary to perform *quantization*, or, perceptual categorization [63]: a track of pitch values is segmented into notes with discrete pitch labels, note timings are quantized to quarter notes, whole notes, and so forth, and timbral information is "quantized" by naming the sound sources involved. In some cases this is not necessary but a parametric or semi-symbolic[4] representation suffices.

An important property of basically all musical cultures is that corresponding notes in different octaves are perceived as having a special kind of similarity, independent of their separation in frequency. The notes C3, C4, and C5, for example, play largely the same harmonic role although they are not interchangeable in a melody. Therefore the set of all notes can be described as representing only 12 *pitch classes*. An individual musical piece usually recruits only a subset of the 12 pitch classes, depending on the *musical key* of the piece. For example, a piece in the C major key tends to prefer the white

[4] In a MIDI file, for example, the time values are not quantized.

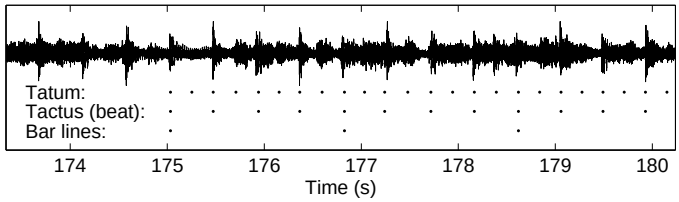


**Fig. 5.** A music signal with three metrical levels illustrated.

keys of the piano, whereas a piece in B major typically employs all the black keys but only two white keys in each octave. Usually there are seven pitch classes that "belong" to a given key. These are called *scale tones* and they possess a varying degree of importance or stability in the key context. The most important is the *tonic* note (for example C in the C major key) and often a musical piece starts or ends on the tonic. Perception of pitch along musical scales and in relation to the musical key of the piece is characteristic to *tonal music*, to which most of Western music belongs [64].

The term *musical meter* has to do with the rhythmic aspects of music: it refers to the regular pattern of strong and weak beats in a piece. Perceiving the meter consists of detecting moments of musical emphasis in an acoustic signal and filtering them so that the underlying periodicities are discovered [65, 62]. The perceived periodicities, *pulses*, at different time scales (or, levels) together constitute the meter, as illustrated in Fig. 5. Perceptually the most salient metrical level is the *tactus* which is often referred to as the foot-tapping rate or the *beat*. The tactus can be viewed as the temporal "backbone" of a piece of music, making beat tracking an important subtask of music transcription. Further metrical analysis aims at identifying the other pulse levels, the periods of which are generally integer multiples or submultiples of the tactus pulse. For example, detecting the *musical measure* pulse consists of determining the number of tactus beats that elapses within one musical measure (usually 2–8) and aligning the boundaries of the musical measures (barlines) to the music signal.

Another element of musical rhythms is *grouping* which refers to the way in which individual sounds are perceived as being grouped into melodic phrases and these are further grouped into larger musical entities in a hierarchical manner [65]. Important to the rhythmic characteristics of a piece of music is how these groups are aligned in time with respect to the metrical system.

The *structure* of a musical work refers to the way in which it can be subdivided into parts and sections at the largest time-scale. In popular music, for example, it is usually possible to identify parts that we label as the chorus, the verse, an introductory section, and so forth. Structural parts can be detected by finding relatively long repeated pitch structures or by observing considerable changes in the instrumentation at section boundaries.

The forthcoming chapters of this book address the extraction and analysis of the above elements in musical audio signals. Fundamental frequency estimation is considered in Parts III and IV of this book, with a separate treatise of melody transcription in Chaps. **??** and **??**. Meter analysis is discussed in Chap. **??** and percussion transcription in Chap. **??**. Chapter **??** discusses the measurement of timbre and musical instrument classification. Structure analysis is addressed in Chap. **??**, and the quantization of time and pitch in Chaps. **??** and **??**, respectively. Before going to a more detailed outline of each chapter, however, let us have a look at some general aspects of the transcription problem.

## 2 Perspectives on Music Transcription

When starting to design a transcription system, certain decisions have to be made already before the actual algorithm development. Among the questions involved are: How should the transcription system be structured into smaller submodules or tasks? What kind of data representations would be the most suitable? Should musical information be used as an aid in the analysis? Would it be advantageous to analyze larger musical structures before going into note-by-note transcription? These general and quite "philosophical" issues are discussed from various perspectives in the following.

### 2.1 Neurophysiological Perspective

First, let us consider a neurophysiological argument into how the music transcription problem should be decomposed into smaller subtasks. In the human auditory cognition, *modularity* of a certain kind has been observed, meaning that certain parts can be functionally and neuro-anatomically isolated from the rest [66, 67, 68]. One source of evidence for this are studies with brain-damaged patients: an accidental brain damage may selectively affect musical abilities but not speech-related abilities, and vice versa [69]. Moreover, there are patients who suffer from difficulties dealing with pitch variations in music but not with temporal variations. In music performance or in perception, either of the two can be selectively lost [70, 66].

Peretz has studied brain-damaged patients who suffer from specific music impairments and she proposes that the music cognition system comprises at least four discernable "modules" [66, 69]. An acoustic analysis module segregates a mixture signal into distinct sound sources and extracts the perceptual parameters of these (including pitch) in some raw form. This is followed by two parallel modules which carry out pitch organization (melodic contour analysis and tonal encoding of pitch) and temporal organization (rhythm and meter analysis). The fourth module, musical lexicon, contains representations of the musical phrases a subject has previously heard.

Neuroimaging experiments in healthy subjects are another way of localizing the cognitive functions in the brain. Speech sounds and higher-level speech information are known to be preferentially processed in the left auditory cortex, whereas musical sounds are preferentially processed in the right auditory cortex [68]. Interestingly, however, when musical tasks involve specifically processing of temporal information (temporal synchrony or duration), the processing is weighted towards the left hemisphere [67], [66]. The relative (not complete) asymmetry between the two hemispheres seems to be related to the acoustic characteristics of the signals: rapid temporal information is characteristic for speech, whereas accurate processing of spectral and pitch information is more important in music [71, 67, 68]. Zatorre et al. proposed that the left auditory cortex is relatively specialized to a better time resolution and the right auditory cortex to a better frequency resolution [67].

In computational transcription systems, rhythm and pitch have often been analyzed separately and using different data representations (see e.g. [18, 28, 72, 19, 73, 20]). Typically, a better time resolution is applied in rhythm analysis and a better frequency resolution in pitch analysis. Based on the above studies, this seems to be justified to some extent. However, it should be kept in mind that studying the human brain is very difficult and the reported results are therefore a subject of controversy. Also, the structure of transcription systems is often determined by merely pragmatic considerations. For example, temporal segmentation is performed prior to pitch analysis in order to allow an appropriate positioning of analysis frames in pitch analysis, which is typically the most demanding stage computationally.

### 2.2 Human Transcription

Another viewpoint to the transcription problem is obtained by studying the conscious transcription process of human musicians and by inquiring about their transcription strategies. The aim of this is to determine the sequence of actions or processing steps that leads to the transcription result.

As already mentioned above, reading and writing music is an acquired ability and therefore the practice of music transcription is of course affected by its teaching at musical institutions. In this context, the term *musical dictation* is used to refer to an exercise where a musical excerpt is played and it has to be written down as notes [74]. An excellent study on the practice of musical dictation and ear training pedagogy can be found in [75].

Characteristic to ear training is that the emphasis is not on trying to *hear* more but to *recognize* what is being heard; to hear relationships accurately and with understanding. Students are presented with different pitch intervals, rhythms, and chords, and they are trained to name these. Simple examples are first presented in isolation and when these become familiar, increasingly complex material is considered. Melodies are typically viewed as a synthesis of pitch and rhythm. For example, Ghezzo instructs the student first to memorize the fragment of music that is to be written down, then to write the pitch of

the notes, and finally to apply the rhythm [74, p.6]. Obviously, ear training presumes a normally-hearing subject who is able to detect distinct sounds and their pitch and timing in the played excerpts – aspects which are very difficult to model computationally.

Recently, Hainsworth conducted a study where he asked trained musicians to describe how they transcribe realistic musical material [76]. The subjects (19 in total) had transcribed music from various genres and with varying goals, but Hainsworth reports that a consistent pattern emerged in the responses. Most musicians first write down the structure of the piece, possibly with some key phrases marked in an approximate way. Next, the chords of the piece or the bass line are notated, and this is followed by the melody. As the last step, the inner lines are studied. Many reported that they heard these by repeated listening, by using an instrument as an aid, or by making musically-educated guesses based on the context.

Hainsworth points out certain characteristics of the above-described process. First, it is sequential rather than concurrent; quoting the author, "no-one transcribes anything but the most simple music in a single pass". In this respect, the process differs from most computational transcription systems. Secondly, the process relies on the human ability to attend to certain parts of a polyphonic signal while selectively ignoring others.[5] Thirdly, some early analysis steps appear to be so trivial for humans that they are not even mentioned. Among these are style detection (causing prior expectations regarding the content), instrument identification, and beat tracking.

### 2.3 Mid-level Data Representations

The concept of *mid-level data representations* provides a convenient way to characterize certain aspects of signal analysis systems. The analysis process can be viewed as a sequence of representations from an acoustic signal towards the analysis result [78, 79]. Usually intermediate abstraction levels are needed between these two since musical notes, for example, are not readily visible in the raw acoustic signal. An appropriate mid-level representation functions as an "interface" for further analysis and facilitates the design of efficient algorithms for this purpose.

The most-often used representation in acoustic signal analysis is the short-time Fourier transform of a signal in successive time frames. Time-frequency decompositions in general are of fundamental importance in signal processing and are introduced in Chap. **??**. Chapter **??** discusses these in a more general framework of *waveform representations* where a music signal is represented as a linear combination of elementary waveforms from a given dictionary. Time-frequency plane representations have been used in many transcription systems (see e.g. [19, 80, 43], and Chap. **??**), and especially in percussive transcription

[5] We may add that also the *limitations* of human memory and attention affect the way in which large amounts of data are written down [77].

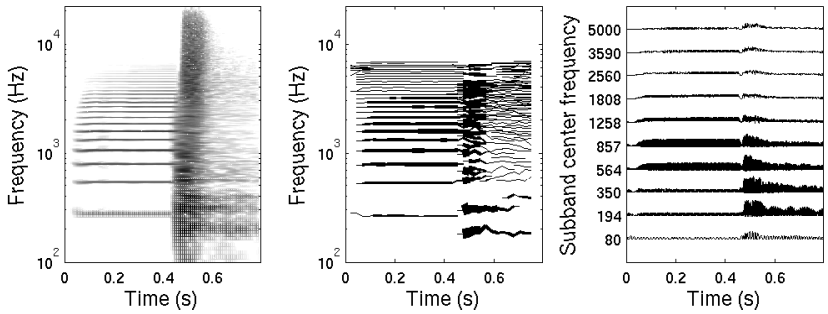


**Fig. 6.** Three different mid-level representations for a short trumpet sound (F0 260 Hz), followed by a snare drum hit. The left panel shows the time-frequency spectrogram with a logarithmic frequency scale. The middle panel shows the sinusoidal model for the same signal, line width indicating the amplitude of each sinusoid. The right panel shows the output of a simple peripheral auditory model for the same signal.

where both linear [45] and logarithmic [46, 25] frequency resolution has been used.

Another common choice for a mid-level representation in music transcription has been the one based on *sinusoid tracks* [18, 81, 35, 82]. In this parametric representation, an acoustic signal is modeled as a sum of sinusoids with time-varying frequencies and amplitudes [83, 84], as illustrated in Fig. 6. Pitched musical instruments can be modeled effectively with relatively few sinusoids and, ideally, the representation supports sound source separation by classifying the sinusoids to their sources. However, this is complicated by the fact that frequency components of co-occurring sounds in music often overlap in time and frequency. Also, reliable extraction of the components in real-world complex music signals can be hard. Sinusoidal models are described in Chap. **??** and applied in Chaps. **??** and **??**.

In the human auditory system, the signal traveling from the inner ear to the brain can be viewed as a mid-level representation. A nice thing about this is that the peripheral parts of hearing are quite well known and computational models exist which are capable of approximating the signal in the auditory nerve to a high accuracy. The right panel of Fig. 6 illustrates this representation. Auditory models have been used for music transcription by several authors [28, 29, 48, 30] and these are further discussed in Chap. **??**.

It is natural to ask if a certain mid-level representation is better than others in a given task. Ellis and Rosenthal have discussed this question in the light of several example representations commonly used in acoustic signal analysis [78]. The authors list several desirable qualities for a mid-level representation. Among these are *component reduction*, meaning that the number of objects in the representation is smaller and the meaningfulness of each is higher compared to the individual samples of the input signal. At the same

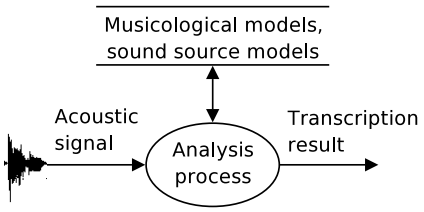


**Fig. 7.** The two main sources of information in music transcription: an acoustic input signal and pre-stored musicological and sound source models.

time, the sound should be decomposed into sufficiently fine-grained elements so as to support *sound source separation* by grouping the elements to their sound sources. Other requirements included *invertibility*, meaning that the original sound can be resynthesized from its representation in a perceptually accurate way, and *psychoacoustic plausibility* of the representation.

### 2.4 Internal Models

Large-vocabulary speech recognition systems are critically dependent on *language models* which represent linguistic knowledge about speech signals [85, 86, 87]. The models can be very primitive in nature, for example merely tabulating the occurrence frequencies of different three-word sequences (N-gram models), or more complex, implementing part-of-speech tagging of words and syntactic inference within sentences.

Musicological information is likely to be equally important for the automatic transcription of polyphonically rich musical material. The probabilities of different notes to occur concurrently or in sequence can be straightforwardly estimated, since large databases of written music exist in an electronic format. Also, there are a lot of musical conventions concerning the arrangement of notes for a certain instrument within a given genre. In principle, these musical constructs can be modeled and learned from data.

In addition to musicological constraints, internal models may contain information about the physics of musical instruments [88], and heuristic rules, for example that a human musician has only ten fingers with limited dimensions. These function as a source of information in the transcription process, along with the input waveform (see Fig. 7). Contrary to an individual music signal, however, these characterize musical tradition at large: its compositional conventions, selection of musical instruments, and so forth. Although these are generally bound to a certain musical tradition, there are also more universal constraints that stem from the human perception (see Bregman [5, Ch. 5]). For example, perceptually coherent melodies usually advance in relatively small pitch transitions and employ a consistent timbre.

Some transcription systems have applied musicological models or sound source models in the analysis [18, 81, 34, 21]. The principles of doing this are

discussed in more detail in Part IV of this book. The term *top-down processing* is often used to characterize systems where models at a high abstraction level impose constraints on the lower levels [89, 90]. In *bottom-up* processing, in turn, information flows from the acoustic signal: features are extracted, combined into sound sources, and these are further processed at higher levels. The "unsupervised-learning" approach mentioned on p. 5 is characterized by bottom-up processing and a minimal use of pre-stored models and assumptions. This approach has a certain appeal too, since music signals are redundant at many levels and, in theory, it might be possible to resolve this "puzzle" in a completely data-driven manner by analyzing a huge collection of musical pieces in connection and by constructing models automatically from the data. For further discussion of this approach, see Chap. **??**.

Utilizing diverse sources of knowledge in the analysis raises the issue of integrating the information meaningfully. In automatic speech recognition, statistical methods have been very successful in this respect: they allow representing uncertain knowledge, learning from examples, and combining diverse types of information.

### 2.5 A Comparison with Speech Recognition

Music transcription is in many ways comparable to automatic speech recognition, although the latter has received greater academic and commercial interest and has been studied longer. Characteristic to both music and speech is that they are *generative* in nature: a limited number of discrete elements are combined to yield larger structures. In speech, phonemes are used to construct words and sentences and, in music, individual sounds are combined to build up melodies, rhythms, and songs. An important difference between the two is that speech is essentially monophonic (one speaker), whereas music is usually polyphonic. On the other hand, speech signals vary more rapidly and the acoustic features that carry speech information are inherently multidimensional, whereas pitch and timing in music are one-dimensional quantities.

A central problem in the development of speech recognition systems is the high dynamic variability of speech sounds in different acoustic and linguistic contexts – even in the case of a single speaker. To model this variability adequately, large databases of carefully annotated speech are collected and used to train statistical models which represent the acoustic characteristics of phonemes and words.

In music transcription, the principal difficulties stem from combinatorics: the sounds of different instruments occur in varying combinations and make up musical pieces. On the other hand, the dynamic variability and complexity of a single sound event is not as high as that of speech sounds. This has the consequence that, to some extent, synthetic music signals can be used in developing and training a music transcriber. Large amounts of training data can be generated since acoustic measurements for isolated musical sounds are

available and combinations of these can be generated by mixing. However, it should be emphasized that this does not remove the need for more realistic acoustic material, too. The issue of obtaining and annotating such databases is discussed in [91] and in Chaps. **??** and **??**. Realistic data are also needed for the objective evaluation of music analysis systems [92].

## 3 Outline

This section discusses the different subtopics of music transcription and summarizes the contents of each chapter of this book. All the chapters are intended to be self-contained entities, and in principle nothing prevents from jumping directly to the beginning of a chapter that is of special interest to the reader. Whenever some element from the other parts of the book is needed, an explicit reference is made to the chapter in question.

### Part I Foundations

The first part of this book is dedicated to topics that are more or less related to all areas of music trancription discussed in this book.

Chapter **??** introduces *statistical and signal processing techniques* that are applied to music transcription in the subsequent chapters. First, the Fourier transform and concepts related to time-frequency representations are described. This is followed by a discussion of statistical methods, including random variables, probability density functions, probabilistic models, and elements of estimation theory. Bayesian estimation methods are separately discussed and numerical computation techniques are described, including Monte Carlo methods. The last section introduces the reader to pattern recognition methods and various concepts related to these. Widely-used techniques such as support vector machines and hidden Markov models are included.

Chapter **??** discusses *sparse adaptive representations for musical signals.* The issue of data representations was already briefly touched in Sect. 2.3 above. This chapter describes parametric representations (for example the sinusoidal model) and "waveform" representations in which a signal is modeled as a linear sum of elementary waveforms chosen from a well-defined dictionary. In particular, signal-adaptive algorithms are discussed which aim at *sparse* representations, meaning that a small subset of waveforms is chosen from a large dictionary so that the sound is represented effectively. This is advantageous from the viewpoint of signal analysis and imposes an implicit structure to the analyzed signal.

### Part II Rhythm and Timbre Analysis

The second part of this book describes methods for meter analysis, percussion transcription, and pitched musical instrument classification.

Chapter **??** discusses *beat tracking and musical meter analysis*, which constitute an important subtask of music transcription. As mentioned on p. 8, meter perception consists of detecting moments of musical stress in an audio signal, and processing these so that the underlying periodicities are discovered. These two steps can also be discerned in the computational methods. Measuring the degree of musical emphasis as a function of time is closely related to *onset detection*, that is, to the detection of the beginnings of discrete sound events in an acoustic signal, a problem which is separately discussed. For the estimation of the underlying metrical pulses, a number of different approaches are described, putting particular emphasis on statistical methods.

Chapter **??** discusses *unpitched percussion transcription*,[6] where the aim is to write down the *timbre class*, or, the sound source, of each constituent sound along with its timing (see Fig. 2 above). The methods discussed in this chapter represent two main approaches. In one, a percussive track is assumed to be performed using a conventional set of drums, such a bass drums, snares, hi-hats, cymbals, tom-toms, and so forth, and the transcription proceeds by detecting distinct sound events and by classifying them into these pre-defined categories. In another approach, no assumptions are made about the employed instrumental sounds, but these are learned from the input signal in an unsupervised manner, along with their occurrence times and gains. This is accomplished by processing a longer portion of the signal in connection and by trying to find such source signals that the percussive track can be effectively represented as a linear mixture of them. Percussion transcription both in the presence and absence of pitched instruments is discussed.

Chapter **??** is concerned with the *classification of pitched musical instrument sounds.* This is useful for music information retrieval purposes, and in music transcription, it is often desirable to assign individual note events into "streams" that can be attributed to a certain instrument. The chapter looks at the acoustics of musical instruments, timbre perception in humans, and basic concepts related to classification in general. A number of acoustic descriptors, or, features, are described that have been found useful in musical instrument classification. Then, different classification methods are described and compared, complementing those described in Chap. **??**. Classifying individual musical sounds in polyphonic music usually requires that they are separated from the mixture signal to some degree. Althoug this is usually seen as a separate task from the actual instrument classification, some methods for the instrument classification in complex music signals are reviewed, too.

## Part III Multiple Fundamental Frequency Analysis

The term *multiple-F0 estimation* refers to the estimation of the F0s of several concurrent sounds in an acoustic signal. The third part of this book describes

[6] Many drum instruments can be tuned and their sound evokes a percept of pitch. Here "unpitched" means that the instruments are not used to play melodies.

different ways to do this. *Harmonic analysis* (writing down the chords of a piece) can be performed based on the results of the multiple-F0 analysis, but this is beyond the scope of this book and an interested reader is referred to [93, Ch.6] and [94, Ch.2]. Harmonic analysis can also be attempted directly, without note-by-note F0 estimation [95, 96, 97].

Chapter **??** discusses *multiple-F0 estimation based on generative models.* Here, the multiple-F0 estimation problem is expressed in terms of a signal model, the parameters of which are being estimated. A particular emphasis in this chapter is laid on statistical methods where the F0s and other relevant parameters are estimated using the acoustic data and possible prior knowledge about the parameter distributions. Various algorithms for online (causal) and offline (non-causal) parameter estimation are described and the computational aspects of the methods are discussed.

Chapter **??** describes *auditory-model based methods for multiple-F0 estimation.* The reader is first introduced with computational models of human pitch perception. Then, transcription systems are described that use an auditory model as a preprocessing step, and the advantages and disadvantages of auditorily-motivated data representations are discussed. The second part of the chapter describes multiple-F0 estimators that are based on an auditory model but make significant modifications to it in order to perform robust F0 estimation in polyphonic music signals. Two different methods are described in more detail and evaluated.

Chapter **??** discusses *unsupervised learning methods for source separation in monaural music signals.* Here the aim is to separate and learn sound sources from polyphonic data without sophisticated modeling of the characteristics of the sources, or detailed modeling of the human auditory perception. Instead, the methods utilize general principles, such as statistical independency between sources, to perform the separation. Various methods are described that are based on independent component analysis, sparse coding, and non-negative matrix factorization.

### Part IV Entire Systems, Acoustic and Musicological Modeling

The fourth part of the book discusses entire music content analysis systems and the use of musicological and sound source models in these.

Chapter **??** is concerned with *auditory scene analysis (ASA) in music signals.* As already mentioned above, ASA refers to the perception of distinct sources in polyphonic signals. In music, ASA aims at extracting entities like notes and chords from an audio signal. The chapter reviews psychophysical findings regarding the acoustic "clues" that humans use to organize spectral components to their respective sources, and the role of internal models and top-down processing in this. Various computational approaches to ASA are described, with a special emphasis on statistical methods and inference in Bayesian networks.

Chapter **??** discusses a research approach called *music scene description*, where the aim is to obtain descriptions that are intuitively meaningful to an untrained listener, without trying to extract every musical note from musical audio. Concretely, this includes the analysis of the melody, bass lines, metrical structure, rhythm, and chorus and phrase repetition. In particular, two research problems are discussed in more detail. *Predominant-F0* estimation refers to the estimation of the F0 of only the most prominent sound in a polyphonic mixture. This closely resembles the experience of an average listener who catches the melody or the "theme" of a piece of music even though he or she would not be able to distinguish the inner lines. Here, methods for extracting the melody and the bass line in music recordings are introduced. The other problem addressed is *music structure analysis*, especially locating the chorus section in popular music.

Chapter **??** addresses *singing transcription*, which means converting a recorded singing performance into a sequence of discrete note pitch labels and their starting and ending points in time. The process can be broken into two stages where, first, a continuous track of pitch estimates (and possibly other acoustic features) is extracted from an acoustic signal, and these are then converted into a symbolic musical notation. The latter stage involves the segmentation of the pitch track into discrete note events and quantizing their pitch values – tasks which are particularly difficult for singing signals. The chapter reviews state-of-the-art singing transcription methods and discusses the use of acoustic and musicological models to tackle the problem.

## References

1. Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Music genre database and musical instrument sound database. In *International Conference on Music Information Retrieval*, pages 229–230, Baltimore, USA, October 2003.
2. The MIDI Manufacturers Association. *The Complete MIDI 1.0 Detailed Specification*, second edition, 1996. Website: `www.midi.org`.
3. E. Selfridge-Field. *Beyond MIDI: the handbook of musical codes*. MIT Press, Cambridge, Massachusetts, 1997.
4. D. Deutsch, editor. *The Psychology of Music*. Academic Press, San Diego, California, 2nd edition, 1999.
5. A.S. Bregman. *Auditory scene analysis*. MIT Press, Cambridge, USA, 1990.
6. J. A. Moorer. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*. PhD thesis, Dept. of Music, Stanford University, 1975. Distributed as Dept. of Music report No. STAN-M-3.
7. J. A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, 1(4):32–38, 1977.
8. C. Chafe, J. Kashima, B. Mont-Reynaud, and J. Smith. Techniques for note identification in polyphonic music. In *International Computer Music Conference*, pages 399–405, Vancouver, Canada, 1985.

9. M. Piszczalski. *A computational model of music transcription*. PhD thesis, Univ. of Michigan, Ann Arbor, 1986.
10. R. C. Maher. *An Approach for the Separation of Voices in Composite Music Signals*. PhD thesis, Univ. of Illinois, Urbana, 1989.
11. R. C. Maher. Evaluation of a method for separating digitized duet signals. *Journal of the Audio Engineering Society*, 38(12):956–979, 1990.
12. Masakata Goto and Yoichi Muraoka. A beat tracking system for acoustic signals of music. In *ACM International Conference on Multimedia*, pages 365–372, San Fransisco, California, October 1994.
13. C. S. Lee. The perception of metrical structure: Experimental evidence and a model. In P. Howell, R. West, and Cross I., editors, *Representing musical structure*. Academic Press, London, 1991.
14. W. Andrew Schloss. *On the automatic transcription of percussive music – from acoustic signal to high-level analysis*. PhD thesis, Center for Computer Research in Music and Acoustics, Stanford University, Stanford, California, USA, May 1985.
15. Jeffrey Adam Bilmes. Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, Massachusetts Institute of Technology, September 1993.
16. Masataka Goto and Yoichi Muraoka. A sound source separation system for percussion instruments. *Transactions of the Institute of Electronics, Information and Communication Engineers D-II*, J77-D-II(5):901–911, May 1994. (in Japanese).
17. A. S. Tanguiane. *Artificial perception and music recognition*. Springer, Berlin Heidelberg, 1993.
18. K. Kashino, K. Nakadai, T. Kinoshita, and H. Tanaka. Organisation of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitative information integration mechanism. In *International Joint Conference on Artificial Intelligence*, pages 158–164, Montreal, Quebec, 1995.
19. M. Goto. A predominant-F0 estimation method for real-world musical audio signals: Map estimation for incorporating prior knowledge about f0s and tone models. In *Proc. Workshop on Consistent and reliable acoustic cues for sound analysis*, Aalborg, Denmark, 2001.
20. M. Davy and S. Godsill. Bayesian harmonic models for musical signal analysis. In *Seventh Valencia International meeting Bayesian statistics 7*, Tenerife, Spain, June 2002.
21. M. Ryynänen and A. Klapuri. Polyphonic music transcription using note event modeling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2005.
22. A. T. Cemgil and B. Kappen. Monte carlo methods for tempo tracking and rhythm quantization. *Journal of Artificial Intelligence Research*, 18:45–81, 2003.
23. S.W. Hainsworth and M.D. Macleod. Particle filtering applied to musical tempo tracking. *Journal of Applied Signal Processing*, 15:2385–2395, 2004.
24. Anssi Klapuri, Antti Eronen, and Jaakko Astola. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Speech and Audio Processing*, 2005. to appear.

25. Olivier Gillet and Gaël Richard. Automatic transcription of drum loops. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004.
26. Jouni K. Paulus and Anssi P. Klapuri. Conventional and periodic N-grams in the transcription of drum sequences. In *IEEE International Conference on Multimedia and Expo*, volume 2, pages 737–740, Baltimore, Maryland, USA, July 2003.
27. P. Herrera, G. Peeters, and S. Dubnov. Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32:3–21, 2003.
28. K. D. Martin. Automatic transcription of simple polyphonic music: robust front end processing. Technical Report 399, MIT Media Laboratory Perceptual Computing Section, 1996.
29. T. Tolonen and M. Karjalainen. A computationally efficient multipitch analysis model. *IEEE Transactions on Speech and Audio Processing*, 8(6):708–716, 2000.
30. A. P. Klapuri. A perceptually motivated multiple-F0 estimation method for polyphonic music signals. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2005.
31. E.D. Scheirer. Tempo and beat analysis of acoustical musical signals. *Journal of the Acoustical Society of America*, 103(1):588–601, January 1998.
32. D. K. Mellinger. *Event Formation and Separation of Musical Sound*. PhD thesis, Stanford University, Stanford, USA, 1991.
33. K. Kashino and H. Tanaka. A sound source separation system with the abilit of automatic tone modeling. In *International Computer Music Conference*, pages 248–255, Tokyo, Japan, 1993.
34. D. Godsmark and G. J. Brown. A blackboard architecture for computational auditory scene analysis. *Speech Communication*, 27(3):351–366, 1999.
35. A. Sterian, M.H. Simoni, and G.H. Wakefield. Model-based musical transcription. In *International Computer Music Conference*, Beijing, China, 1999.
36. A. Hyväinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.
37. Michael Anthony Casey. *Auditory Group Theory with Applications to Statistical Basis Methods for Structured Audio*. PhD thesis, Massachusetts Institute of Technology, 1998.
38. M. A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *International Computer Music Conference*, Berlin, Germany, 2000.
39. Philippe Lepain. Polyphonic pitch extraction from musical signals. *Journal of New Music Research*, 28(4):296–309, 1999.
40. Paris Smaragdis. *Redundancy Reduction for Computational Audition, a Unifying Approach*. PhD thesis, Massachusetts Institute of Technology, 1997.
41. Paris Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2003.
42. S. A. Abdallah. *Towards Music Perception by Redundancy Reduction and Unsupervised Learning in Probabilistic Models*. PhD thesis, Department of Electronic Engineering, King's College London, 2002.
43. S. A. Abdallah and M. D. Plumbley. Polyphonic transcription by non-negative sparse coding of power spectra. In *International Conference on Music Information Retrieval*, pages 318–325, Barcelona, Spain, October 2004.

44. Derry FitzGerald. *Automatic Drum Transcription and Source Separation*. PhD thesis, Dublin Institute of Technology, 2004.
45. D. FitzGerald, E. Coyle, and B. Lawlor. Prior subspace analysis for drum transcription. In *Audio Engineering Society 114th Convention*, Amsterdam, Netherlands, March 2003.
46. Jouni Paulus and Tuomas Virtanen. Drum transcription with non-negative spectrogram factorisation. In *European Signal Processing Conference*, Antalya, Turkey, September 2005.
47. H. Kameoka, T. Nishimoto, and S. Sagayama. Separation of harmonic structures based on tied gaussian mixture model and information criterion for concurrent sounds. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, 2004.
48. M. Marolt. A connectionist approach to transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6(3):439–449, 2004. URL: lgm.fri.uni-lj.si/SONIC.
49. Aymeric Zils, Francois Pachet, Olivier Delerue, and Fabien Gouyon. Automatic extraction of drum tracks from polyphonic music signals. In *International Conference on Web Delivering of Music*, Darmstadt, Germany, December 2002.
50. Derry FitzGerald, Bob Lawlor, and Eugene Coyle. Drum transcription in the presence of pitched instruments using prior subspace analysis. In *Irish Signals & Systems Conference 2003*, Limerick, Ireland, July 2003.
51. Kazuyoshi Yoshii, Masataka Goto, and Hiroshi G. Okuno. Drum sound identification for polyphonic music using template adaptation and matching methods. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, October 2004.
52. F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Speech and Audio Processing*, 2005. to appear.
53. K. Kashino and H. Murase. A sound source identification system for ensemble music based on template adaptation and music stream extraction. *Speech Communication*, 27:337–349, 1999.
54. A. L. Berenzweig and D. P. W. Ellis. Locating singing voice segments within music signals. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 119–122, New Paltz, USA, October 2001.
55. J. Eggink and G. J. Brown. Instrument recognition in accompanied sonatas and concertos. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 217–220, Montreal, Canada, 2004.
56. E. Vincent and X. Rodet. Instrument identification in solo and ensemble music using independent subspace analysis. In *International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.
57. Thomas D. Rossing. *The Science of Sound*. Addison Wesley, second edition, 1990.
58. W. M. Hartmann. Pitch, periodicity, and auditory organization. *Journal of the Acoustical Society of America*, 100(6):3491–3502, 1996.
59. C. J. Plack and Carlyon R. P. Loudness perception and intensity coding. In Moore [60], pages 123–160.
60. B. C. J. Moore, editor. *Hearing—Handbook of Perception and Cognition*. Academic Press, San Diego, California, 2nd edition, 1995.
61. S. Handel. Timbre perception and auditory object identification. In Moore [60], pages 425–460.

62. E. F. Clarke. Rhythm and timing in music. In Deutsch [4], pages 473–500.
63. E. M. Burns. Intervals, scales, and tuning. In Deutsch [4], pages 215–264.
64. C. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, 1990.
65. F. Lerdahl and R. Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, 1983.
66. I. Peretz. Music perception and recognition. In B. Rapp, editor, *The Handbook of Cognitive Neuropsychology*, pages 519–540. Hove: Psychology Press, 2001.
67. R. J. Zatorre, P. Belin, and V. B. Penhune. Structure and function of auditory cortex: music and speech. *TRENDS in Cognitive Sciences*, 6(1):37–46, 2002.
68. M. Tervaniemi and K. Hugdahl. Lateralization of auditory-cortex functions. *Brain Research Reviews*, 43(3):231–46, 2003.
69. I. Peretz and M. Coltheart. Modularity of music processing. *Nature Neuroscience*, 6(7), 2003.
70. S. D. Bella and I. Peretz. Music agnosias: Selective impairments of music recognition after brain damage. *Journal of New Music Research*, 28(3):209–216, 1999.
71. Robert V. Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995.
72. M. Goto and Y. Muraoka. Music understanding at the beat level: real-time beat tracking for audio signals. In *International Joint Conference on Artificial Intelligence*, pages 68–75, Montreal, Quebec, 1995.
73. A. Klapuri, T. Virtanen, A. Eronen, and J. Seppänen. Automatic transcription of musical recordings. In *Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, Aalborg, Denmark, 2001.
74. M. A. Ghezzo. *Music theory, ear training, rhythm, solfége, and dictation, a comprehensive course*. The University of Alabama Press, Alabama, 1980.
75. D. P. Hedges. *Taking notes: The history, practice, and innovation of musical dictation in English and American aural skills pedagogy*. PhD thesis, School of Music, Indiana University, 1999.
76. S.W. Hainsworth. *Techniques for the Automated Analysis of Musical Audio*. PhD thesis, Department of Engineering, University of Cambridge, 2004.
77. B. Snyder. *Music and memory*. MIT Press, Cambridge, Massachusetts, 2000.
78. D. P. W. Ellis and D. F. Rosenthal. Mid-level representations for computational auditory scene analysis. In *International Joint Conference on Artificial Intelligence*, Montreal, Quebec, 1995.
79. D. Marr. *Vision*. W. H. Freeman and Company, 1982.
80. A. P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Speech and Audio Processing*, 11(6):804–815, 2003.
81. K. D. Martin. A blackboard system for automatic transcription of simple polyphonic music. Technical Report 385, MIT Media Laboratory Perceptual Computing Section, 1996.
82. T. Virtanen and A. Klapuri. Separation of harmonic sound sources using sinusoidal modeling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 765–768, Istanbul, Turkey, 2000.
83. R.J. McAulay and Th.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:744–754, 1986.

84. X. Serra. Musical sound modeling with sinusoids plus noise. In C. Roads, S. Pope, A. Picialli, and G. De Poli, editors, *Musical Signal Processing*. Swets & Zeitlinger, 1997.
85. L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, 1993.
86. F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, Massachusetts, 1997.
87. Daniel Jurafsky and James H. Martin. *Speech and language processing*. Prentice Hall, New Jersey, USA, 2000.
88. N.H. Fletcher and T.D. Rossing. *The Physics of Musical Instruments*. Springer, Berlin, Germany, second edition, 1998.
89. M. Slaney. A critique of pure audition. In *International Joint Conference on Artificial Intelligence*, pages 13–18, Montreal, Quebec, 1995.
90. D. P. W. Ellis. Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis, and its application to speech/nonspeech mixtures. *Speech Communication*, 27:281–298, 1999.
91. M. Lesaffre, M. Leman, B. De Baets, and J.-P. Martens. Methodological considerations concerning manual annotation of musical audio in function of algorithm development. In *International Conference on Music Information Retrieval*, Barcelona, Spain, 2004.
92. MIREX: Annual Music Information Retrieval Evaluation eXchange, 2005. URL: www.music-ir.org/ mirexwiki/ index.php/ MIREX_2005.
93. D. Temperley. *The cognition of basic musical structures*. MIT Press, Cambridge, Massachusetts, 2001.
94. R. Rowe. *Machine musicianship*. MIT Press, Cambridge, Massachusetts, 2001.
95. M. Leman. *Music and schema theory*. Springer, Heidelberg, 1995.
96. F. Carreras, M. Leman, and M. Lesaffre. Automatic description of musical signals using schema-based chord decomposition. *Journal of New Music Research*, 28(4):310–331, 1999.
97. Alexander Sheh and Daniel P.W. Ellis. Chord segmentation and recognition using EM-trained hidden Markov models. In *International Conference on Music Information Retrieval*, pages 183–189, Baltimore, USA, October 2003.

# Index