Journal of New Music Research
2004, Vol. 33, No. 3, pp. 269–282

Routledge
Taylor & Francis Group

Automatic Music Transcription as We Know it Today

Anssi P. Klapuri

Tampere University of Technology, Tampere, Finland

Abstract

The aim of this overview is to describe methods for the automatic transcription of Western polyphonic music. The transcription task is here understood as transforming an acoustic musical signal into a MIDI-like symbolic representation. Only pitched musical instruments are considered: recognizing the sounds of drum instruments is not discussed. The main emphasis is laid on estimating the multiple fundamental frequencies of several concurrent sounds. Various approaches to solve this problem are discussed, including methods that are based on modelling the human auditory periphery, methods that mimic the human auditory scene analysis function, signal model-based Bayesian inference methods, and data-adaptive methods. Another subproblem addressed is the rhythmic parsing of acoustic musical signals. From the transcription point of view, this amounts to the temporal segmentation of music signals at different time scales. The relationship between the two subproblems and the general structure of the transcription problem is discussed.

1. Introduction

Transcription of music is here defined as the process of analyzing an acoustic musical signal so as to write down the musical parameters of the sounds that occur in it. Traditionally, written music uses *note symbols* to indicate the pitch, onset time, and duration of each sound to be played. Loudness and applied musical instruments are not specified for individual notes but are determined for larger parts.

In a representational sense, music transcription can be seen as transforming an acoustic signal into a symbolic representation. However, written music is primarily a *performance instruction*, rather than a representation of music. It describes music in a language that a musician understands and can use to produce musical sound. From this point of view, music transcription can be viewed as discovering the "recipe", or, reverse-engineering the "source code" of a music signal. The applied notation does not necessarily need to be the traditional musical notation but any symbolic representation is adequate if it gives sufficient information for performing a piece using the available musical instruments. In the case that an electronic synthesizer is used for resynthesis, a MIDI file is an example of an appropriate representation.

A musical notation does not only allow reproducing a piece of music but also modifying, rearranging, and processing music at a high abstraction level. Another important application of music transcription is *structured audio coding*. A MIDI-like representation is extremely compact yet retains the identifiability and characteristics of a piece of music to an important degree. Even when sound source parameters are included, the required bandwidth remains around 2–3 kbit/s (ISO, 1999). Other uses of music transcription comprise *information retrieval* based on, e.g., the melody of a piece, *musicological analysis* of improvised and ethnic music, and *interactive music systems* which generate an accompaniment to the singing or playing of a soloist (Raphael, 2001a,b; Rowe, 2001).

Automatic transcription of polyphonic[1] music has been the subject of increasing research interest during the last 10 years. The problem is in many ways analogous to that of automatic speech recognition but has not received comparable academic or commercial interest. Longer-term research projects have been undertaken at Stanford University (Moorer, 1975; Chafe et al., 1986), University of Michigan

[1] Here *polyphonic music* refers to a signal where several sounds occur simultaneously. In *monophonic* signals, at most one note is sounding at a time.

Correspondence: Anssi P. Klapuri, Tampere University of Technology, Institute of Signal Processing, Korkeakoulunkatu 1, FIN-33100 Tampere, Finland. Tel.: +358 50 364 8208; Fax: +358 3 3115 4954; E-mail: klap@cs.tut.fi

DOI: 10.1080/0929821042000317840

(Piszczalski et al., 1979, 1986; Sterian, 1999), University of Tokyo (Kashino et al., 1993, 1995), Massachusetts Institute of Technology (Hawley, 1993; Martin, 1996a,b), University of London (Bello, 2003; Abdallah et al., submitted), Cambridge University (Hainsworth, 2001, 2003), and at Tampere University of Technology (Klapuri, 1998; Eronen, 2001; Paulus et al., 2003; Viitaniemi et al., 2003; Virtanen, 2003). Doctoral theses on the topic have been written by Moorer (1975), Watson (1985), Piszczalski (1986), Maher (1989), Mellinger (1991), Hawley (1993), Godsmark (1998), Rossi (1998), Sterian (1999), Marolt (2002), Bello (2003), Hainsworth (2003), and Klapuri (2004).

Despite the number of attempts to solve the problem, a practically applicable, general-purpose transcription system does not exist at the present time. The proposed systems fall clearly behind skilled human musicians in accuracy and flexibility. The most recent proposals, however, have achieved a certain degree of accuracy in transcribing polyphonic music of limited complexity (Kashino et al., 1995; Martin, 1996b; Sterian, 1999; Tolonen et al., 2000; Davy et al., 2003; Bello, 2003). The typical constraints on the target signals are that the number of concurrent sounds is limited (or fixed), and that inter ference of drums and percussive instruments is often not allowed. Some degree of success for real-world music on CD recordings has been previously demonstrated by Goto (2001), who attempted to extract the melody and the bass lines from music signals. Also, estimating the number of concurrent sounds and suppressing percussive instruments has been addressed by Klapuri (2003).

Surprisingly, even the transcription of single-voice singing is not a solved problem, as indicated by the fact that the accuracy of the "voice input" functionalities in score-typesetting programs is still very limited. *Tracking the pitch* of a monophonic musical passage is practically a solved problem, but *quantization* of the continuous track of pitch estimates into note symbols with discrete pitch and durations has turned out to be very difficult for some signals, particularly for singing (Viitaniemi et al., 2003). A comparative evaluation of the available monophonic transcribers can be found in (Clarisse et al., 2002).

1.1 Terminology

Some terms have to be clarified before going any further. *Pitch* is a perceptual attribute of sounds, defined as the frequency of a sine wave that is matched to the target sound in a psychoacoustic experiment (Hartmann, 1996). If the matching cannot be accomplished consistently by human listeners, the sound does not have a pitch. *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. In ambiguous situations, the period corresponding to the perceived pitch is chosen. The term *multiple-F0 estimation* is here used to refer to the estimation of the F0s of several concurrent sounds.

The term *musical meter* refers to the regular pattern of strong and weak beats in a piece of music. Metrical analysis, here also called *rhythmic parsing*, refers to the process of detecting moments of musical stress in an acoustic signal and filtering them so that the underlying periodicities are discovered (Lerdahl et al., 1983). The perceived periodicities (*pulses*) at different time scales together constitute the meter. Metrical analysis at a certain time scale is taking place for example when a person taps his foot to music.

1.2 Decomposition of the music transcription problem

It is useful to decompose the music transcription problem into smaller and more approachable subproblems. First of all, multiple-F0 analysis and rhythmic parsing have often been performed separately and using different data representations (Kashino, 1995; Martin, 1996b; Goto, 1995, 2000; Davy, 2003). Based on what we know about the modularity of music processing in the human brain, this seems to be justified and not only a technical artefact (Bella et al., 1999; Peretz, 2001). Typically, a better time resolution is applied in meter analysis and a better frequency resolution in F0 analysis.

Note that rhythmic parsing and multiple-F0 estimation are complementary to each other. An accurate multiple-F0 estimator is able to indicate which notes are active at each time but is often not able to decide the exact beginning or end times of discrete note events. Metrical analysis, in turn, generates a temporal framework which can be used to divide the music signal into meaningful temporal segments and to quantize the timing of musical events.

Another efficient way of structuring the transcription problem is through so-called *mid-level representations*. Auditory perception may be viewed as a hierarchy of representations from an acoustic signal up to a conscious percept (Ellis, 1995). Usually intermediate abstraction level(s) are needed since note symbols, for example, cannot be easily observed in the raw acoustic signal as such. A well-defined mid-level representation functions as an "interface" for the higher-level inference that follows. A fundamental mid-level representation in our hearing is the signal in the auditory nerve. Whereas we know rather little about the exact mechanisms of the brain, rather accurate *auditory models* exist which are able to approximate the signal in the auditory nerve (Moore, 1995). This is a great advantage, since an important part of auditory analysis takes place already at the peripheral stages.

Finally, there are two main sources of information to be used in music transcription. Pre-stored internal models concerning the structure of music signals constitute a source of information in addition to the incoming acoustic waveform. Large-vocabulary speech recognition systems are critically dependent on language models that represent linguistic knowledge concerning speech signals (Jurafsky, 2000). Musicological information is equally important for the automatic transcription of polyphonically rich musical material.

The probabilities of particular notes occurring simultaneously or sequentially can be straightforwardly estimated, since large databases of written music exist in an electronic format. More complex rules governing music are available in the theory of music and composition. Temperley has proposed a very comprehensive rule-based system to analyze the musical structure of MIDI files (Temperley, 2001). From the transcription point of view, a remaining challenge is to transform these rule-based models into probabilistic models which can be used to evaluate the likelihoods of several candidate analyses already during the transcription process.

1.3 Scope and purpose of this article

The aim of this overview is to describe methods for the automatic transcription of Western polyphonic music. Only pitched musical instruments are considered; detecting or recognizing the sounds of percussive (drum) instruments is not discussed. The interested reader is referred to (Gouyon et al., 2001; FitzGerald et al., 2002; Zils et al., 2002; Paulus et al., 2003). Also, emphasis is laid on the acoustic signal analysis part and not so much on musicological models. The main part of this paper is dedicated to what is considered to be the core of the music transcription problem, multiple-F0 estimation.

This paper is organized as follows. Section 2 will describe properties of musical sounds and the basic principles of F0 estimation. In Section 3, different approaches to multiple-F0 estimation will be introduced. Work on rhythmic parsing will be discussed in Section 4 and future prospects of music transcription in Section 5.

2. Musical sounds and F0 estimation

The human auditory system tries to assign a pitch to almost all kinds of acoustic signals (Meddis, 1991). In the case of F0 estimation algorithms, the scope has to be limited to periodic or nearly periodic sounds for which the concept of F0 is defined. For many algorithms, the target signals are further limited to so-called *harmonic sounds*. These are discussed next.

2.1 Harmonic sounds versus non-harmonic sounds

Harmonic sounds have a spectral structure where the significant frequency components are regularly spaced. Figure 1 illustrates a harmonic sound in the time and frequency domains.

For an ideal harmonic sound, the frequencies of the overtone partials (harmonics) are integer multiples of the F0. In the case of many real-world sound production mechanisms, however, the partial frequencies are not in *exact* integral ratios although the general structure of the spectrum is similar to that in Figure 1. For real strings, for example, the frequencies of the partials obey the formula

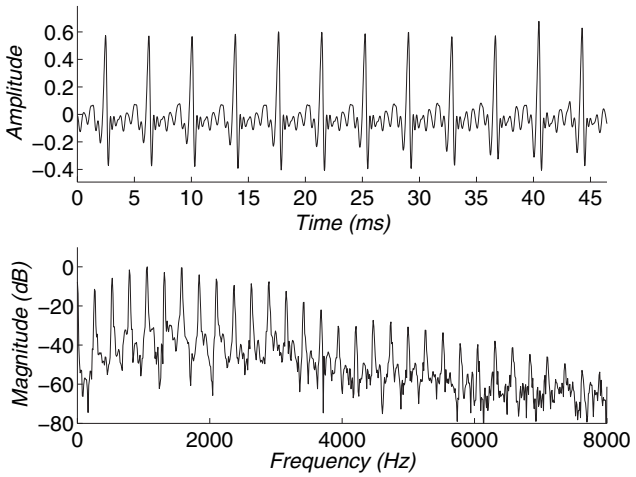

Fig. 1. A harmonic sound illustrated in the time and frequency domains. The example represents a trumpet sound with fundamental frequency 260 Hz and fundamental period 3.8 ms.

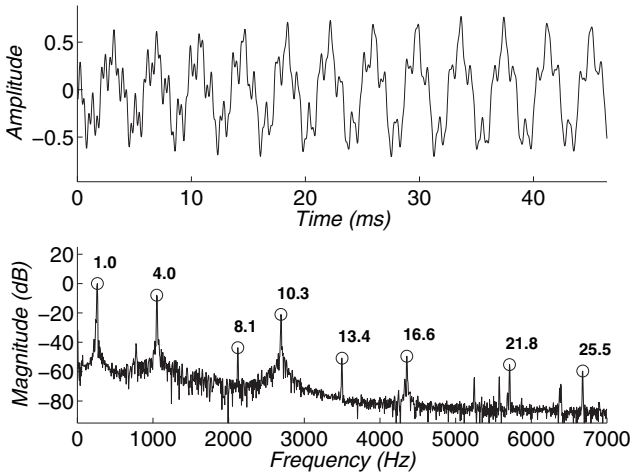

Fig. 2. A vibraphone sound with fundamental frequency 260 Hz illustrated in the time and frequency domains. In the lower panel, frequencies of the most dominant spectral components are shown in relation to the F0.

$$f_h = hF\sqrt{1+\beta\left(h^2-1\right)} \tag{1}$$

where F is the fundamental frequency, h the harmonic index (partial number), and β the inharmonicity factor (Fletcher et al., 1998). The *inharmonicity* phenomenon is due to the stiffness of real strings and causes the higher-order partials to be slightly shifted upwards in frequency. However, the structure of the spectrum is in general very similar to that in Figure 1 and the sounds belong to the class of harmonic sounds.

Figure 2 shows an example of a sound which does not belong to the class of harmonic sounds although it is nearly periodic in the time domain and has a clear pitch. In Western music, *mallet percussion instruments* are a case in point; these instruments produce pitched sounds which are not harmonic.

Many of the F0 estimation methods discussed in the following are primarily concerned with harmonic sounds. However, this limitation is not very severe in Western music where harmonic sounds prevail. Harmonic sounds are produced by string instruments, reed instruments, brasses, flutes, pipe organs, and by the human voice. Non-harmonic sounds are produced by mallet percussions (marimba, vibraphone, xylophone, glockenspiel) and drums.

There is no single obvious way of calculating the F0 of a signal that is not perfectly periodic. The F0 estimation algorithms do not only differ in technical details, but also in regard to the very information that the calculations are based on. In psychoacoustics, computational models of pitch perception have been traditionally classified as either place models or temporal models. A good introduction to these competing theories and their supporting evidence can be found in (Hartmann, 1996). In the case of practical F0 estimation methods, a different categorization is more useful. Algorithms are here grouped into those that look for frequency partials at harmonic *spectral locations* and those that observe *spectral intervals* (frequency intervals) between partials. The underlying idea of both of these approaches can be understood by looking at Figure 1. Algorithms which measure periodicity of the time-domain signal belong to the first category. Algorithms which measure periodicity of the Fourier spectrum belong to the latter category. Algorithms which measure periodicity of the time-domain amplitude envelope represent a tradeoff between the two classes.

2.2 Spectral-location type F0 estimators

Time-domain autocorrelation function (ACF) based algorithms are among the most frequently used F0 estimators (see, e.g., Brown et al., 1991; Talkin, 1995). Even very recently, a novel and very accurate ACF-oriented F0 estimator has been proposed by de Cheveigné and Kawahara (2002). Usability of the latter method for music transcription has been evaluated by Viitaniemi et al. (2003) and Klapuri (2003).

As pointed out by Tolonen and Karjalainen (2002), ACF-based F0 estimators have close similarities on the model level with cepstrum-based F0 estimators (Noll, 1967), and there is a continuum between the two. This becomes evident when calculating the autocorrelation function $r(\tau)$ of a time-domain signal $x(n)$ via the discrete Fourier transform (DFT) and its inverse (IDFT) as

$$r(\tau) = \mathrm{IDFT}\left(\left|\mathrm{DFT}(x(n))\right|^2\right). \tag{2}$$

Definition of the cepstrum $c(\tau)$ of $x(n)$ is analogous to Equation (2) and is obtained by replacing the second power with a logarithm function. Tolonen et al. have suggested combining the advantages of the two by using a "generalized autocorrelation function" where the second power is replaced with a real-valued exponent (0.67 in their case) (Tolonen et al., 2000).

Both ACF and cepstrum-based F0 detectors are implicit realizations of a model which emphasizes frequency partials at *harmonic locations* of the magnitude spectrum. This can be seen by writing Equation (2) in terms of the Fourier spectrum $X(k)$ of a real-valued input signal as

$$r(\tau) = \frac{1}{K}\sum_{k=0}^{K-1}\left[\cos\left(\frac{2\pi\tau k}{K}\right)|X(k)|^2\right], \tag{3}$$

where K is the length of the transform frame. The function $\cos(2\pi\tau k/K)$ assigns unity weights to spectral components at integer multiples of a F0 candidate τ/f_s. Here f_s is the sampling rate.

Another way of weighting frequency components according to their spectral locations is to perform harmonic pattern-matching in the frequency domain. Different strategies for doing this have been proposed, e.g., by Brown (1992), Doval et al. (1993), and Maher et al. (1994).

A shortcoming of the spectral-location oriented F0 estimators is that they do not take the inharmonicity phenomenon into account. As mentioned, the higher partials of many real musical instruments cannot be assumed to reside at exactly harmonic spectrum positions.

2.3 Spectral-interval type F0 estimators

The spectrum-autocorrelation method and its variants have been successfully used in several F0 estimators (see, e.g., Lahat et al., 1987; Kunieda et al., 1996). The idea is derived from the observation that harmonic sounds have a periodic magnitude spectrum, the period of which is the F0. In its simplest form, the autocorrelation function $\tilde{r}(m)$ over the positive frequencies of a K-length magnitude spectrum can be calculated as

$$\tilde{r}(m) = \frac{2}{K}\sum_{k=0}^{K/2-m-1}|X(k)||X(k+m)|, \tag{4}$$

Here, any two spectral components with frequency interval m support the corresponding F0 candidate mf_s/K. The spectrum can be arbitrarily shifted without affecting the output value. Building F0 calculations upon the *intervals* between frequency partials works better for sounds that exhibit inhar monicities. Even though the intervals do not remain constant, they are more stable than the locations of the partials, which shift cumulatively, as can be seen in Equation (1).

Another interesting difference between spectral-location and spectral-interval based approaches is that the former methods are prone to errors in F0 halving and the latter to errors in F0 doubling. Consider the conventional time-domain ACF estimator; the signal is periodic at twice the fundamental period as well, and this easily leads to F0 *halving*. In the case of the frequency-domain ACF, the magnitude spectrum is periodic at *double* the F0 rate but shows no periodicity at half the F0 rate.

2.4 Periodicity of the time-domain amplitude envelope

Above, algorithms were discussed which measure the periodicity of the time-domain signal or the periodicity of the Fourier spectrum. A third, fundamentally different approach is to measure the periodicity of the time-domain amplitude envelope. This approach has been widely used in systems that attempt to model the human auditory system (Meddis, 1991, 1997; Tolonen, 2000; Klapuri, 2004). The underlying idea is that any signal with more than one frequency component exhibits periodic fluctuations, *beating*, in its time-domain amplitude envelope. That is, the partials alternatingly amplify and cancel each other, depending on their phase. The rate of the beating caused by each pair of frequency components depends on their frequency difference. In the case of a harmonic sound, the frequency interval corresponding to the F0 dominates.

Envelope periodicity is typically analyzed for several subbands of an input signal. Figure 3 illustrates the beating phenomenon for the harmonic overtones 15–19 of a signal with 220 Hz fundamental frequency. The amplitude envelope of the signal is obtained by half-wave rectifying and lowpass filter ing the signal in the time domain. The fundamental period (4.5 ms) is clearly visible in the resulting time-domain signal. The half-wave rectification (HWR) operation is defined as

$$\mathrm{HWR}(x(n)) = \max(x(n), 0) \tag{5}$$

As can be seen in Figure 3(b), the spectrum of the amplitude envelope around zero frequency consists of beating components which correspond to frequency intervals between the partials in the input signal.

Performing the HWR operation for several bandpass-filtered versions of a signal allows an elegant synthesis of the spectral-location and spectral-interval based approaches. Note that prior to the lowpass filtering the rectified signal contains both the partials of the input signal *and* the beating components which represent frequency intervals between the input partials. This becomes evident when the nonlinear HWR operation is approximated in the frequency domain. Let $x(n)$ be a zero-mean gaussian random process and $X(k)$ its complex Fourier spectrum. As described in (Klapuri, 2004), the complex Fourier spectrum $Y(k)$ of the half-wave rectified signal $y(n)$ = HWR($x(n)$) can be quite accurately approximated by

$$\hat{Y}(k) = \frac{\sigma_x}{\sqrt{8\pi}}\delta(k) + \frac{1}{2}X(k) + \frac{1}{\sigma_x\sqrt{8\pi}}\sum_{j=-K/2+k}^{K/2-k} X(j)X(k-j), \tag{6}$$

where $\delta(k)$ is the unit impulse function and σ_x is the standard deviation of $x(n)$. On the right-hand side of Equation (6), the first term is a dc-component, the second term represents the spectrum of the input signal, and the last term, convolution of the spectrum $X(k)$ with itself, represents the beating components of the amplitude-envelope spectrum. In addition, the last term generates a harmonic distortion spectrum centered on twice the center frequency of the input narrowband signal $x(n)$.

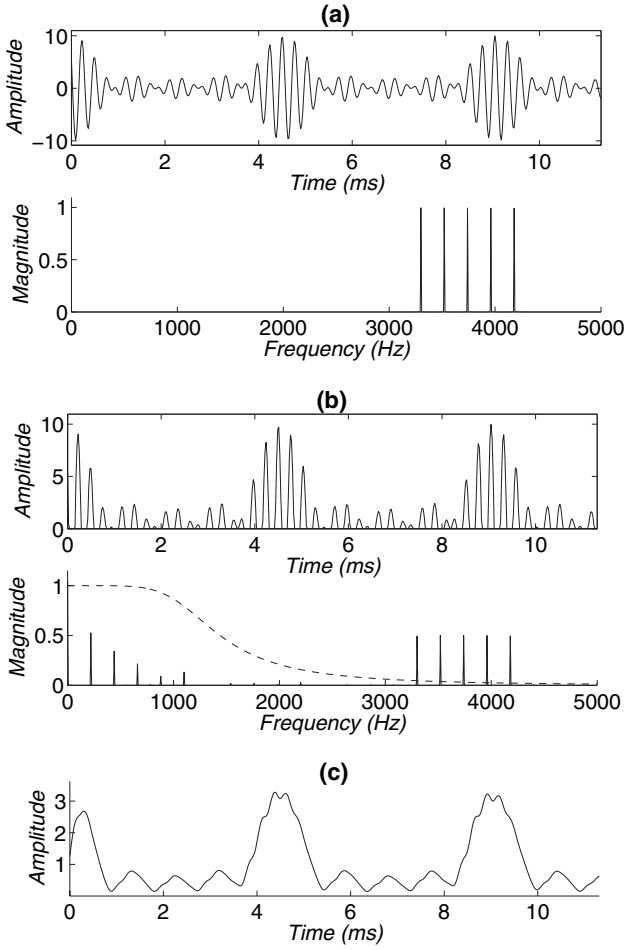

Fig. 3. (a) A signal containing partials 15–19 of a sound with F0 = 220 Hz. (b) The signal after half-wave rectification. (c) The signal after rectification and lowpass filtering. The response of the lowpass filter is shown as a dashed line in (b).

If the cutoff frequency of the lowpass filter in Figure 3 is raised so that the filter passes the spectral components of the original narrowband signal in an appropriate proportion, a subsequent periodicity analysis (ACF computation, for example) utilizes both spectral-location and spectral-interval information. This leads to a more reliable F0 analysis (Klapuri, 2004). In Section 3.2, some transcription systems will be introduced that are based on analyzing the periodicity of the amplitude envelope in subbands.

3. Multiple-F0 estimation

The aim of multiple-F0 estimation is to find the F0s of all the component sounds in a mixture of signals. The complexity of this task is significantly higher than that of monophonic F0 estimation. Some intuition of the difference in complexity can be developed by comparing the spectrum of

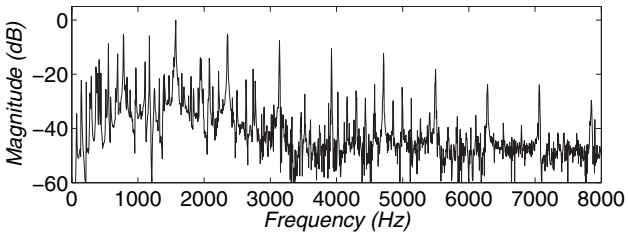

Fig. 4. The spectrum of a mixture of four harmonic sounds.

a mixture of four harmonic sounds in Figure 4 with that of one sound in Figure 1.

The diversity of approaches taken towards multiple-F0 estimation is even wider than that in single-F0 estimation. Also, it is difficult to categorize multiple-F0 estimation methods according to any single taxonomy because the methods are complex and typically combine several processing principles. As a consequence, there is no single dimension which could function as an appropriate basis for categorization. However, some main viewpoints to the problem can be discerned and it is the aim of this section to introduce these.

In the following, a few representative examples from each of the main viewpoints to the problem are described in more detail. The given list of methods is not intended to be complete. For a comprehensive historical overview of F0 estimation methods in music, see Hainsworth (2001). Subheadings are pro vided to improve readability but it should be remembered that the cited papers really cannot be put under a single label.

3.1 Auditory-scene-analysis oriented approach

Multiple-F0 estimation is closely related to sound separation. An algorithm that is able to estimate the F0s of several concurrent sounds is, in effect, also organizing the respective spectral components to their sound sources (Bregman, 1990, p. 240). Vice versa, if an accurate separation algorithm was available, a conventional F0 estimation method could be used to measure the F0s of the separated sounds.

The human auditory system is very effective in perceiving and recognizing individual sound sources in mixture signals. This cognitive function is called *auditory scene analysis* (ASA). Computational modeling of ASA has been the subject of increasing research interest since 1990 when Bregman (1990) published a comprehensive description of the principles and mechanisms of the ASA in humans. Recent overviews of computational ASA (CASA) can be found in (Ellis, 1996; Rosenthal et al., 1998; Cooke et al., 2001).

CASA is usually viewed as a two-stage process where an incoming signal is first *decomposed* into its elementary time-frequency components, and these are then *organized* to their respective sound sources. Bregman pointed out a number of measurable acoustic "cues" which promote the grouping of

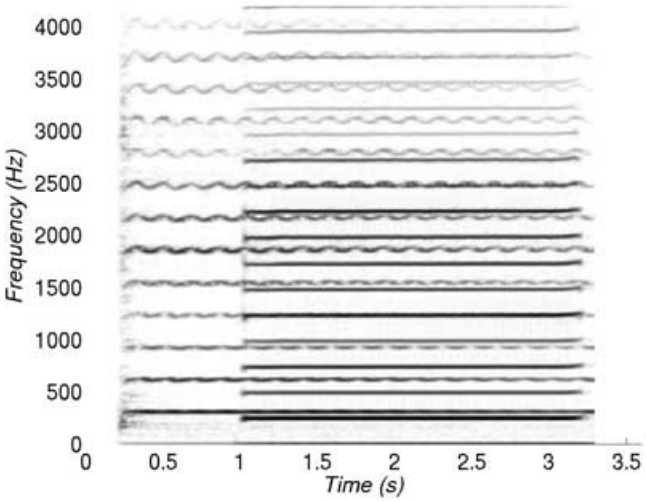

Fig. 5. Time-frequency representation of a mixture of two harmonic sounds: a cello sound (F0 = 310 Hz) which starts at 250 ms and a saxophone sound (F0 = 250 Hz) which starts at 1.0 s.

time-frequency components to the same sound source in human listeners. Among these are: proximity in time-frequency, harmonic frequency relationships, synchronous changes in the frequency or amplitude of the components, and spatial proximity (i.e., the same direction of arrival).

Figure 5 shows the spectrogram of a mixture of two harmonic sounds. Many of the above-mentioned cues are visible in the figure. The partials of the cello sound start 750 ms before the saxophone sound and exhibit synchronous frequency modulation. Also, the partials within each sound are in harmonic relationships. Such features are commonly present in acoustic signals and facilitate the perceptual separation of the component sounds.

Time-varying sinusoidal components, *sinusoidal tracks*, have often been used as the elementary components for which the mentioned features are measured (Kashino, 1993, 1995; Sterian, 1999). Extraction of these components from music signals has been addressed, e.g., in Serra (1997) and Levine (1998). More recently, also auditorily-motivated time-frequency representations have been used (Godsmark et al., 1999).

Kashino et al. (1993, 1995) brought Bregman's ideas to music scene analysis and also proposed several other new ideas for music transcription. In their system, sinusoidal tracks were clustered into note hypotheses by applying a subset of the perceptual grouping cues mentioned above. Harmonicity rules and onset timing rules were implemented. In addition, timbre models were used to identify the musical instrument of each note, and pre-stored tone memories for each instruments were used to resolve colliding frequency components. Chordal analysis was performed based on the probabilities of different notes to occur under a given chord. Chord transition probabilities were encoded into trigram models (Markov chains). For computations, a Bayesian probability network was used to integrate the processing mecha-

nisms and to do simultaneous bottom–up analysis, temporal tying, and top–down processing (chords predict notes and notes predict spectral components). Evaluation material comprised five different instruments and polyphonies of up to three simultaneous sounds. Later, Kashino et al. (1999) addressed the problem of source identification and source stream formation when the F0 information is given.

A doctoral thesis by Sterian (1999) was more tightly focused on implementing the perceptual grouping principles for the purpose of music transcription. Sinusoidal partials were used as the mid-level representation. Sterian implemented the perceptual grouping rules as a set of likelihood functions, each of which evaluated the likelihood of the observed partials given a hypothesized grouping of the partials to note candidates. Distinct likelihood functions were defined to take into account onset and offset timing, harmonicity, low partial support, partial gap, and partial density (see the reference for the definitions of the latter concepts). The product of all these likelihood functions was used as the criterion for optimal grouping. While an exhaustive search over all possible groupings was not possible, a multiple-hypothesis tracking strategy was used to find a suboptimal solution.

Godsmark and Brown (1999) used a computational model of the peripheral auditory system to extract "synchrony strands" (dominant time-frequency components in different bands) for which the grouping cues were extracted. The "blackboard" architecture applied in their system was particularly designed to facilitate the integration of, and competition between, the perceptual grouping principles. The strands were organized to sound events and these were further grouped to their respective sources (event "streams") by computing pitch and timbre proximities between successive sound events. The model was evaluated by show ing that it could segregate melodic lines from polyphonic music. Transcription accuracy as such was not the main goal.

3.2 Methods based on modeling the auditory periphery

The "unitary" pitch model of Meddis et al. (1991, 1997) has had a strong influence on the F0 estimation research. While Bregman's theory is primarily concerned with the psychology of auditory perception, the unitary model addresses the more peripheral (largely physiological) parts of hearing. Simulation experiments with the model have shown that this single model (hence the name unitary model) is capable of reproducing a wide range of phenomena in human pitch perception.

The unitary model analyzes the amplitude-envelope periodicity at the outputs of a bank of bandpass filters. It consists of the following processing steps (Meddis et al., 1997).

1. an acoustic input signal is passed though a bank of 40–120 bandpass filters;
2. the signal in each channel is compressed, half-wave rectified, and lowpass filtered;
3. periodicity estimation within channels is carried out by calculating short-time ACF estimates;
4. the ACF estimates are summed across channels to obtain a *summary autocorrelation function*:

$$s(\tau) = \sum_c r_c(\tau), \tag{7}$$

where $r_c(\tau)$ is the autocorrelation function at subband c. The maximum value of $s(\tau)$ is then used to indicate the perceived pitch period.

Cheveigné and Kawahara (1999) extended the unitary model to multipitch estimation in mixture signals. They proposed a system where pitch estimation was followed by cancellation of the detected sound, and the estimation was then repeated for the residual signal. The cancellation was performed either by subband selection or by performing within-band cancellation filtering. Although evaluation results were reported only for synthetic and perfectly periodic signals, the proposed iterative approach was indeed a successful one.

Tolonen and Karjalainen (2000) developed a computationally efficient version of the unitary pitch model and applied it to the multiple-F0 estimation of musical sounds. Only two subbands were used instead of the 40–120 bands in the original model, yet the main characteristics of the model were preserved. Practical robustness was addressed by flattening the spectrum of an incoming sound by inverse warped-linear-prediction filtering, and by using the generalized ACF (see Section 2.2) for periodicity estimation. Extension to multiple-F0 estimation was achieved by cancelling subharmonics in the output of the model. From the resulting *enhanced summary autocorrelation function*, all F0s were picked without iterative estimation and cancellation. The method is relatively accurate and has been statistically evaluated (Klapuri, 2003).

Martin used the log-lag correlogram model of Ellis (1996) as an auditory front-end in his system (Martin, 1996a,b). The blackboard architecture was applied to integrate knowledge about physical sound production, rules governing tonal music, and "garbage collection" heuristics. Support for different F0s was raised on a frame-by-frame basis and then combined with longer-term power-envelope information to create note hypotheses. Musical rules favoured F0s in certain intervallic relations. Transcription results were shown for a few example cases.

Klapuri (2004) proposed certain modifications to the unitary pitch model in order to obtain a reliable multiple-F0 estimation tool for use in music signals. The first two steps of the unitary model (bandpass filtering, compression, and rectification) were retained but the ACF calculations were replaced by a technique called *harmonic selection* and a more complex subband-weighting was applied when combining the results across bands. Computational efficiency was achieved by approximating the HWR operation in the frequency domain according to Equation (6). The method was evaluated by calculating error rates for random mixtures of

recorded musical instrument samples. Good accuracy was achieved using analysis frame sizes of 46 ms or longer.

3.3 Signal-model based probabilistic inference

Musical signals are highly structured. It is possible to state the whole multiple-F0 estimation prob lem in terms of a signal model, the parameters of which should be estimated. Consider, e.g., the model (Davy et al., 2003):

$$y(t) = \sum_{n=1}^{N} \sum_{m=1}^{M_n} [a_{n,m} \cos(m\omega_n t) + b_{n,m} \sin(m\omega_n t)] + e(t) \quad (8)$$

where N is the number of simultaneous sounds, M_n is the number of partials in sound n, ω_n is the fundamental frequency of sound n, and $a_{n,m}$, $b_{n,m}$ together encode the amplitude and phase of individual partials. The term $e(t)$ is a residual noise component.

In principle, all the parameters on the right-hand side of the above equation should be estimated based on the observation $y(t)$ and possible prior knowledge about the parameter distributions. As pointed out by Davy et al. (2003), the problem is Bayesian in the sense that there is a lot of prior knowledge concerning music signals.

Davy and Godsill (2003) elaborated the above signal model to accommodate time-varying amplitudes, non-ideal harmonicity, and non-white residual noise. A likelihood function for observing $y(t)$ given model parameters was defined. Prior distributions for the parameters were carefully selected. An input signal was first segmented into excerpts where no note transitions occur. Then the parameters of the signal model were estimated in the *time domain*, separately for each segment. The main problem of this approach is in the actual computations. For any sufficiently realistic signal model, the parameter space is huge and the posterior distribution is highly multimodal and strongly peaked. Davy and Godsill used variable-dimension Markov chain Monte Carlo sampling of the posterior, reporting that much of the innovative work was spent on finding heuristics for the fast exploration of the parameter space. Although computationally inefficient, the system was reported to work quite robustly for polyphonies up to three simultaneous sounds.

Goto (2001) has proposed a method which models the *short-time spectrum* of a music signal as a weighted mixture of tone models. Each tone model consists of a fixed number of harmonic components which are modeled as Gaussian distributions centered on integer multiples of the F0 in the spectrum. Goto derived a computationally feasible expectation-maximization (EM) algorithm which iteratively updates the tone models and their weights, leading to maximum *a posteriori* parameter estimates. Temporal continuity was considered by tracking framewise F0 weights within a multiple-agent architecture. Goto used the algorithm successfully to track the melody and the bass lines in real-time on CD recordings. Although the overall system of Goto is relatively complex, the core EM algorithm can be easily implement based on the reference. The algorithm estimates the weights of all F0s, but typically only one (predominant) F0 was found in our simulations, exactly as claimed by Goto.

3.4 Data-adaptive techniques

In data-adaptive systems, there is no parametric model or other knowledge of the sources. Instead, the source signals are estimated from the data. Typically, it is not even assumed that the sources (which here refer to individual notes) have harmonic spectra. For real-world signals, the performance of, e.g., independent component analysis alone is poor. However, by placing certain restrictions on the sources, the data-adaptive techniques become applicable in realistic cases. Such restrictions are, e.g., independence of the sources and *sparseness* which means that the sources are assumed to be inactive most of the time.

Virtanen (2003) added *temporal continuity* constraint to the sparse coding paradigm. He used the signal model

$$S(t, f) = \sum_{n=1}^{N} a_{t,n} S_n(f) + S_e(t, f), \quad (9)$$

where the power spectrogram of the input, $S(t, f)$ is represented as a linear sum of N static source spectra $S_n(f)$ with time-varying gains $a_{t,n}$. The term $S_e(t, f)$ represents the error spectrogram. Virtanen proposed an iterative optimization algorithm which estimates non-negative $a_{t,n}$ and $S_n(f)$ based on the minimization of a cost function which takes into account reconstruction error, sparseness, and temporal continuity. The algorithm was used to separate pitched and drum instruments in real-world music signals (Virtanen, 2003).

Also Abdallah et al. (submitted) applied sparse coding for the analysis of music signals. Input data was represented as magnitude spectrograms, and sources as magnitude spectra, leading to a source mixing model which is essentially the same as in Equation (9). The authors proposed an algorithm where sources were obtained using gradient-ascent inference and the time-varying gains with maximum-likelihood learning. Their results were promising, although shown only for one example case, a synthesized Bach piece with two to three simultaneous sounds.

3.5 Other approaches

An important line of research has been pursued by Okuno, Nakatani, and colleaques who have demonstrated effective use of the direction-of-arrival information in segregating simultaneous speakers (Nakatani et al., 1999; Okuno et al., 1999). The system of Nakatani et al. was designed to segregate continuous streams of harmonic sounds, such as the voiced sections of two or three simultaneous speakers. Multiple agents were deployed to trace harmonic sounds in stereo signals. Detected sounds were cancelled from the input signal and the residual was used to update the parameters of each sound and to detect new sounds. Wang and Brown

(1999) have used F0 information for speech segregation. A multi-pitch tracking algorithm for noisy speech was presented by Wu et al. (2002).

The periodicity transform method of Sethares and Staley (1999) is an example of a mathematical viewpoint to multiple-F0 estimation. The authors proposed a residue-driven sound separation algorithm where one periodic component at a time was estimated and cancelled from the mixture signal. The overall iterative approach resembles that of de Cheveigné et al. (1999).

Marolt (2001) has used neural networks for the different subproblems of music transcription. The author proposed a system which is a combination of an auditory model, adaptive oscillators, and neural networks. The unitary pitch model (see Section 3.2) was used to process an input signal and adaptive oscillators, similar to those in Large et al. (1994), were then used to track partials at the output of each frequency channel. In order to track harmonically related partials, the oscillators were connected to oscillator nets, one per candidate musical note. A distinct neural network was trained for each individual note in order to detect its presence.

A potentially very successful approach in some applications is to focus on modeling a specific musical instrument. This has been done, e.g., in Hawley (1993), Rossi (1998) and Klapuri (1998) where only piano music was considered.

4. Rhythmic parsing

Rhythmic parsing is an essential part of understanding music signals. From the point of view of music transcription, metrical analysis amounts to *temporal segmentation* of music according to certain criteria. As already mentioned in Section 1.2, multiple-F0 estimation and metrical analysis complement each other. Imagine a time-frequency plane where time flows from left to right and different F0s are arranged in an ascending order on the vertical axis. On top of this plane, a multiple-F0 estimator produces horizontal lines which indicate the probabilities of different notes to be active as a function of time. Metrical analysis, in turn, produces a framework of vertical "grid lines" which can be used to segment the note activation curves into discrete note events and to quantize their timing.

Automatic meter analysis as such has several applications. To name a few examples, the resulting temporal framework facilitates cut-and-paste operations and editing of music signals, and enables synchronization with light effects or video.

Moments of musical stress, *accents*, are important for metrical analysis. These serve as cues from which a human listener extrapolates a regular metrical pulse (Lerdahl et al., 1983). Accentuated events in music are communicated by the beginnings of all discrete sound events, especially the onsets of long pitched events, sudden changes in loudness or timbre, and harmonic changes.

Musical meter is a hierarchical structure, consisting of pulse sensations at different time scales, or, levels. The most prominent level is the *tactus*, often referred to as the foot-tapping rate or the beat. *Tempo* of a piece is defined as the rate of the tactus pulse. In a musically meaningful meter, the period lengths of larger-scale metrical pulses are integer multiples of the periods at lower levels (Lerdahl et al., 1983).

4.1 Early work

The earliest computational models of meter analysis were designed to process symbolic data (impulse patterns or musical scores). An extensive comparison of the early models can be found in (Lee, 1991; Desain et al., 1999). In brief, the models can be seen as being based on a *set of rules* that were used to define what makes a musical accent and to infer the most natural meter. The rule system proposed by Lerdahl and Jackendoff (1983) is the most complete, but was described in verbal terms only.

Allen et al. (1990) proposed a method to track the tactus pulse of MIDI performances in real-time. The authors used beam search to track multiple competing interpretations. Rosenthal (1992) proposed a system to emulate the human rhythm perception for piano performances, pre sented as MIDI files. In his system, other auditory organization functions were modeled, too, grouping notes into streams and chords. Parncutt (1994) proposed a detailed algorithmic meter perception model based on systematic listening tests.

A more straightforward signal-processing oriented approach was taken by Brown (1993) who analyzed the meter of musical scores using the autocorrelation function. Large and Kolen (1994) proposed to estimate meter using adaptive oscillators which adjust their period and phase to an incoming pattern of impulses which were located at the onsets of musical events.

4.2 More recent meter estimation systems

Table 1 lists characteristic attributes of more recent meter analysis systems. The systems can be classified into two main categories according to whether they process MIDI files or acoustic signals. The column "evaluation material" gives a more specific idea of the musical material that each system has been tested on. Another defining characteristic is the output of different systems. Many algorithms analyze the meter only at the tactus pulse level. Some others produce useful side-information, for example quantizing the onset times of musical events. The column "technique" attempts to summarize the method that is used to achieve the analysis result.

As a part of a larger project of modeling the cognition of basic musical structures, Temperley and Sleator (1999) proposed a meter analysis algorithm for arbitrary MIDI files. The algorithm is based on implementing the preference rules verbally described in Lerdahl et al. (1983) and produces the

Table 1. Characteristics of some meter estimation systems.

Reference	Input	Output	Technique	Evaluation material
Temperley & Sleator (1999)	MIDI	Meter, time quantization	Rule-based approach; implementation of the preference rules in (Lerdahl et al.,1983)	Example analyses; all music types; source code available
Dixon (2001)	MIDI, audio	Tactus	First find periods using IOI histogram, then phases using multiple agents	222 MIDI files, 10 audio files; source codeavailable
Raphael (2001)	MIDI, audio	Tactus, time quantization	Probabilistic generative model for onset times; MAP estimation (Viterbi)	Two example analyses; expressive performances
Cemgil & Kappen (2003)	MIDI	Tactus, time quantization	Probabilistic generative model for onset times; sequential Monte Carlo methods	216 polyphonic piano performances of 12 Beatles songs; clave pattern
Goto & Mur-aoka (1995, 1997)	Audio	Meter	Extract onset components; IOI histogram; multiple tracking agents	85 pieces; pop music with and without drums, 4/4 time signature
Scheirer (1998)	Audio	Tactus	Bank of comb filters to analyze periodicity of power envelopes at six subbands	60 pieces with "strong beat"; all music types; source code available
Laroche (2001)	Audio	Tactus, swing	Extract discrete onsets; maximum-likelihood estimation	Qualitative report; music with constant tempo and sharp attacks
Sethares & Staley (2001)	Audio	Meter	Calculate RMS-energies at 1/3-octave subbands; apply a periodicity transform	A few examples; music with constant tempo
Gouyon et al. (2002)	Audio	Tatum	First find periods (IOI histogram), then phases by matching isochronous pattern	57 drum sequences, each 2–10 s. in duration; constant tempo
Klapuri et al. (to appear)	Audio	Meter	Measure degree of accentuation; bank of comb filters; probabilistic model	474 audio signals; all music types

whole metrical hierarchy as the output. Dixon (2001) proposed a rule-based system to track the tactus pulse of expressive MIDI performances. Also, he introduced a simple onset detector to make the system applicable for audio signals. The source codes of both Temperley's and Dixon's system are publicly available for testing.

Cemgil and Kappen (2003) developed a *probabilistic generative model* for the event times in expressive musical performances. They used the model to infer a hidden continuous tempo variable and quantized ideal note onset times from observed noisy onset times in a MIDI file. Tempo tracking and time quantization were performed simultaneously so as to balance the smoothness of tempo deviations versus the complexity of the resulting quantized score. A similar Bayesian model was independently proposed by Raphael (2001a,b), who also demonstrated its use for acoustic input.

Goto and Muraoka (1995, 1997) were the first to present a meter-tracking system which works with a reasonable accuracy for audio signals. Their system operated in real time and was based on an architecture where multiple agents tracked alternative metrical hypotheses. Beat positions at the larger levels were inferred by detecting certain drum sounds (Goto & Muraoka, 1995) or chord changes (Goto & Muraoka, 1997). Gouyon et al. (2002) proposed a system for estimating the temporally atomic "tatum" pulse in percussive audio tracks of constant tempo. The authors computed an inter-onset interval histogram and applied the two-way mismatch method of Maher (1994) to find the tatum ("temporal atom") which best explained multiple harmonic peaks in the histogram. Laroche (2001) used a straightforward probabilistic model to estimate the tempo and swing[2] of audio signals. Input to the model was provided by a relatively simple onset detector.

Scheirer (1998) proposed a method for tracking the tactus pulse of music signals of any kind, provided that they had a "strong beat". Important in Scheirer's approach was that he did not detect discrete onsets or sound events as a middle-step, but performed periodicity analysis directly on the half-wave rectified differentials of subband power envelopes. Periodicity in each subband was analyzed using a bank of comb-filter resonators. The source code of Scheirer's system is publicly available for testing. Sethares and Staley (2001) took a similar approach but used a periodicity transform for periodicity analysis instead of a bank of comb filters.

Klapuri et al. (to appear) proposed a system for estimating the meter of acoustic music signals at three time scales: the tactus, tatum, and musical measure pulse levels. For the initial time-frequency analysis, a technique was proposed which measures the degree of musical accentuation as a function of time in four different frequency ranges. Periodicity analysis of the accent signals was performed using a bank of comb filter resonators similar to those used by Scheirer (1998). The comb filters served as feature extractors for a

[2] *Swing* is a characteristic of musical rhythms most commonly found in jazz. The term refers to a non-isochronous pulse consisting of alternating long and short inter-beats intervals.

probabilistic model where the tatum, tactus, and measure pulses were jointly estimated. The probabilistic model encoded prior musical knowledge regarding well-formed musical meters by taking into account dependencies between the three pulse levels and by implementing temporal tying between successive metrical hypotheses. Evaluation results were shown for a database of music comprising all the main Western genres.

4.3 Summary and discussion

Automatic meter analysis in general can be performed more accurately than multiple-F0 estimation. For this reason, many transcription systems perform temporal segmentation prior to F0 analysis and use the metrical information for positioning the analysis frames in the latter part. Similarly to human listeners, computational meter analysis seems to be easiest at the tactus pulse level which is usually the most prominent level aurally (Lerdahl et al., 1983).

There are a few basic problems that a rhythmic parser has to address to be successful. First, the degree of musical accentuation as a function of time has to be measured. In the case of audio input, this has much to do with the initial time-frequency analysis and is closely related to the problem of sound onset detection. Some systems measure accentuation in a continuous manner (Scheirer, 1998; Sethares et al., 2001; Klapuri et al., to appear), whereas others extract discrete events (Goto et al., 1995, 1997; Laroche, 2001; Gouyon, 2002). Robustness for diverse signal types, for example classical versus rock music, represents a challenge.

Secondly, the periods and phases of the underlying metrical pulses have to be estimated. The methods which detect discrete events as a middle step have often used inter-onset interval histograms for estimating the period (Goto et al., 1995, 1997; Dixon, 2001; Gouyon, 2002). An internal musical model of some kind is needed to achieve stable meter tracking and to fill in parts where the meter is only faintly implied by the musical surface.

Thirdly, a system has to choose the metrical level which corresponds to the tactus or some other specially designated pulse level. This may take place implicitly, or by using a prior distribution for pulse periods (Parncutt, 1994), or by applying rhythmic pattern matching (Goto et al., 1995), to name a few examples. In some cases, choosing the metrical level is ambiguous even for a human listener. Typically, however, at least the tactus pulse can be unambiguously determined. Tempo halving or doubling in a meter analysis system is a symptom of failing in the latter task.

5. When will music transcription be a "solved problem"?

An important fact about music transcription is that it is *difficult*. The problem is at best comparable to automatic speech recognition which has been studied for 50 years and is only now becoming practically applicable. In music transcription, the development will probably be faster as the computational power is already available and we can borrow theoretical methods and approaches from speech recognition. However, the problem is really not in finding fast computers but in discovering the mechanisms and principles that humans use when listening to music. Modelling perception is difficult because the world in which we live is complex and because the human brain is complex. It is very unlikely that there would be a quick solution to the polyphonic transcription problem, or a single mechanism that would solve the problem once and for all. The human brain combines a large number of processing principles and heuristics. We will be searching for them for years, perhaps even decades, before arriving at a skilled musician's accuracy and flexibility.

There is a certain factor which may crucially change the above prediction regarding the *time* needed to release an accurate music transcriber. This has to do with the generative nature of music versus speech. The development of speech recognition systems is constantly confronted with the problem that the amount of targeted and carefully annotated training data is limited. In music transcription, the very problem stems from *combinatorics*: the sounds of different instruments occur in varying combinations and make up musical pieces. The dynamic variability and complexity of a single sound event is not as high as that of speech sounds.[3] For this reason, *synthetic music*, to a certain degree, is valid for training a music transcriber. Very large amounts of training data can be generated since acoustic measurements for isolated musical sounds are available, combinations of these can be generated by mixing, and effects can be added. An exact reference annotation is immediately available.

The availability of training data helps us to automate the most frustrating part of algorithm development: parameter optimization. However, it does not free us from designing the methods them selves. It is quite unlikely that the transcription problem could be solved simply by training a huge neural network, for example. The "space" of possible algorithms and models may be even larger than we can think of. The interesting part is to explore this space in a meaningful and efficient way until we have found the necessary ingredients of a successful transcription system.

References

Abdallah, S.A., & Plumbley, M.D. (Submitted). Sparse coding of music signals. *Neural Computation.*

Allen, P.E., & Dannenberg, R.B. (1990). Tracking musical beats in real time. In: *Proc. 1990 International Computer Music Conference*, 140–143.

Bella, S.D., & Peretz, I. (1999). Music agnosias: Selective impairments of music recognition after brain damage. *Journal of New Music Research*, *28*, 209–216.

[3] Even for singing, transcribing the melody is significantly easier than recognizing the lyrics.

Bello, J.P. (2003). Towards the automated analysis of simple polyphonic music: A knowledge-based approach. Ph.D. thesis, University of London.

Bregman, A.S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Brown, J.C., & Zhang, B. (1991). Musical frequency tracking using the methods of conventional and "narrowed" autocorrelation. *Journal of the Acoustical Society of America*, *89*, 2346–2354.

Brown, J.C. (1992). Musical fundamental frequency tracking using a pattern recognition method. *Journal of the Acoustical Society of America*, *92*, 1394–1402.

Brown, J.C. (1993). Determination of the meter of musical scores by autocorrelation. *Journal of the Acoustical Society of America*, *94*, 1953–1957.

Cemgil, A.T., & Kappen, B. (2003). Monte Carlo methods for tempo tracking and rhythm quantization. *Journal of Artificial Intelligence Research*, *18*, 45–81.

Chafe, C., & Jaffe, D. (1986). Source separation and note identification in polyphonic music. In: *Proc. IEEE International Conf. on Acoust., Speech, and Signal Processing*, Tokyo, 1289–1292.

Clarisse, L.P., Martens, J.P., Lesaffre, M., De Baets, B., De Mayer, H., & Leman, M. (2002). An auditory model based transcriber of singing sequences. In: *Proc. International Conference on Music Information Retrieval*, Paris, France.

Cooke, M., & Ellis, D.P.W. (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, *35*, 141–177.

Davy, M., & Godsill, S.J. (2003). Bayesian harmonic models for musical signal analysis. In: J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian Statistics VII*, Oxford: Oxford University Press.

de Cheveigné, A., & Kawahara, H. (1999). Multiple period estimation and pitch perception model. *Speech Communication*, *27*, 175–185.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, *111*, 1917–1930.

Desain, P., & Honing, H. (1999). Computational models of beat induction: the rule-based approach. *Music Perception*, *11*, 29–42.

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, *30*, 39–58.

Doval, B., & Rodet, X. (1993). Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMM's. In: *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 221–224.

Ellis, D.P.W., & Rosenthal, D.F. (1995). Mid-level representations for computational auditory scene analysis. In: *Proc. Workshop on Computational Auditory Scene Analysis*, *Int. Joint Conf. on Artif. Intell.*, Montreal.

Ellis, D.P.W. (1996). *Prediction-driven computational auditory scene analysis*. Ph.D. thesis. Cambridge, MA: MIT Media Laboratory.

Eronen, A. (2001). Comparison of features for musical instrument recognition. In: *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York.

FitzGerald, D., Coyle, E., & Lawlor, B. (2002). Sub-band independent subspace analysis for drum transcription. In: *Proc. 5th Int. Conference on Digital Audio Effects (DAFX-02)*, Hamburg, Germany, 65–69.

Fletcher N.F., & Rossing, T.D. (1998). *The Physics of Musical Instruments.* 2nd edn. New York: Springer.

Godsmark, D. (1998). A computational model of the perceptual organization of polyphonic music. PhD thesis, University of Sheffield.

Godsmark, D., & Brown, G.J. (1999). A blackboard architecture for computational auditory scene analysis. *Speech Communication*, *27*, 351–366.

Goto, M., & Muraoka, Y. (1995). Music understanding at the beat level – real-time beat tracking for audio signals. In: *Working Notes of the IJCAI-95 Workshop on Computational Auditory Scene Analysis*, 68–75.

Goto, M., & Muraoka, Y. (1997). Real-time rhythm tracking for drumless audio signals – chord change detection for musical decisions. In: *Proc. IJCAI-97 Workshop on Computational Auditory Scene Analysis*, 135–144.

Goto, M. (2001). 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.

Gouyon, F., & Herrera, P. (2001). Exploration of techniques for automatic labeling of audio drum tracks' instruments. In: *Proc. MOSART: Workshop on Current Directions in Computer Music*, Barce lona, Spain.

Gouyon, F., Herrera, P., & Cano, P. (2002). Pulse-dependent analyses of percussive music. In: *Proc. AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, 396–401.

Hainsworth, S.W. (2001). *Analysis of musical audio for polyphonic transcription*. 1st year report, Department of Engineering, University of Cambridge.

Hainsworth, S.W. (2003). *Techniques for the automated analysis of musical audio*. Ph.D. thesis, Cambridge University, UK.

Hartmann, W.M. (1996). Pitch, periodicity, and auditory organization. *Journal of the Acoustical Society of America*, *100*, 3491–3502.

Hawley, M. (1993). *Structure out of sound*. Ph.D. thesis. Cambridge, MA: MIT Media Laboratory.

International Organization for Standardization. (1999). *Information technology – coding of audiovis ual objects; Part 3: audio; Subpart 5: structured audio.* ISO/IEC FDIS 14496-3 sec5. [On-line]. Available: http://web.media.mit.edu/~eds/mpeg4/SA-FDIS.pdf.

Jurafsky, D., & Martin, J.H. (2000). *Speech and Language Processing*. Englewood Cliffs, NJ: Prentice Hall.

Kashino, K., & Tanaka, H. (1993). A sound source separation system with the ability of automatic tone modeling. In:

Proc. International Computer Music Conference, Tokyo, 248–255.

Kashino, K., Nakadai, K., Kinoshita, T., & Tanaka, H. (1995). Organization of hierarchical perceptual sounds: music scene analysis with autonomous processing modules and a quantitative information integration mechanism. In: *Proc. International Joint Conf. on Artificial Intelligence*, Montréal.

Kashino, K., & Murase, H. (1999). A sound source identification system for ensemble music based on template adaptation and music stream extraction. *Speech Communication*, *27*, 337–349.

Klapuri, A.P. (1998). *Automatic transcription of music*. MSc thesis, Tampere University of Technology, Finland.

Klapuri, A.P. (2003). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. In: *IEEE Trans. Speech and Audio Proc.*, December 2003.

Klapuri, A.P. (2004). *Signal processing methods for the automatic transcription of music*. Ph.D. thesis, Tampere University of Technology, Finland.

Klapuri, A.P., Eronen J.E., & Astola J.T. (in press). Analysis of the meter of acoustic musical signals. *IEEE Trans. Speech and Audio Proc*.

Kunieda, N., Shimamura, T., & Suzuki, J. (1996). Robust method of measurement of fundamental frequency by ACLOS – autocorrelation of log spectrum. In: *Proc. IEEE International Conf. on Acoust., Speech, and Signal Processing*, 232–235.

Lahat, M., Niederjohn, R.J., & Krubsack, D.A. (1987). A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. *IEEE Trans. Acoust., Speech, and Signal Processing*, *35*, 741–750.

Large, E.W., & Kolen, J.F. (1994). Resonance and the perception of musical meter. *Connection Science*, *6*, 177–208.

Laroche, J. (2001). Estimating tempo, swing and beat locations in audio recordings. In: *Proc. Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, 135–138.

Lee, C.S. (1991). The perception of metrical structure: experimental evidence and a model. In: P. Howell, R. West, & I. Cross (Eds.), *Representing Musical Structure*. London: Academic Press.

Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Levine, S.N. (1998). Audio representation for data compression and compressed domain processing. Ph.D. thesis, University of Stanford.

Maher, R.C. (1989). An approach for the separation of voices in composite music signals. Ph.D. thesis, University of Illinois, Urbana.

Maher, R.C., & Beauchamp, J.W. (1994). Fundamental frequency estimation of musical signals using a two-way mismatch procedure. *Journal of the Acoustical Society of America*, *95*, 2254–2263.

Marolt, M. (2001). SONIC: Transcription of polyphonic piano music with neural networks. In: *Proc. Workshop on Current Research Directions in Computer Music*, Barcelona.

Marolt, M. (2002). *Transcription of polyphonic solo piano music*. Ph.D. thesis, University of Ljubljana, Slovenia.

Martin, K.D. (1996a). A blackboard system for automatic transcription of simple polyphonic music. MIT Media Laboratory Perceptual Computing Section Technical Report No. 385.

Martin, K.D. (1996b). Automatic transcription of simple polyphonic music: robust front end processing. MIT Media Laboratory Perceptual Computing Section Technical Report No. 399.

Meddis, R., & Hewitt, M.J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America*, *89*, 2866–2882.

Meddis, R., & O'Mard, L. (1997). A unitary model of pitch perception. *Journal of the Acoustical Society of America*, *102*, 1811–1820.

Mellinger, D.K. (1991). *Event formation and separation in musical sound*. Ph.D. thesis, Center for Computer Research in Music and Acoustics, Stanford University.

Moore, B.C.J. (Ed.) (1995). *Hearing – Handbook of Perception and Cognition*, 2nd edn. San Diego, CA: Academic Press.

Moorer, J.A. (1975). On the segmentation and analysis of continuous musical sound by digital computer. Ph.D. thesis, Department of Music, Stanford University.

Nakatani, T., & Okuno, H.G. (1999). Harmonic sound stream segregation using localization and its application to speech stream segregation. *Speech Communication*, *27*, 209–222.

Noll, A.M. (1967). Cepstrum pitch detection. *Journal of the Acoustical Society of America*, *41*, 293–309.

Okuno, H.G., Nakatani, T., & Kawabata, T. (1999). Listening to two simultaneous speeches. *Speech Communication*, *27*, 299–310.

Parncutt, R. (1994). A perceptual model of pulse salience and metrical accent in musical rhythms. *Music Perception*, *11*, 409–464.

Paulus, J.K., & Klapuri, A.P. (2003). Conventional and periodic N-grams in the transcription of drum sequences. In: *Proc. IEEE International Conferences on Multimedia and Expo*, Baltimore, MD, USA.

Peretz, I. (2001). Music perception and recognition. In: B. Rapp (Ed.), *The Handbook of Cognitive Neuropsychology*, Hove: Psychology Press, 519–540.

Piszczalski, M., & Galler, B.A. (1979). Predicting musical pitch from component frequency ratios. *Journal of the Acoustical Society of America*, *66*, 710–720.

Piszczalski, M. (1986). *A computational model of music transcription*. PhD thesis, University of Michigan, Ann Arbor.

Raphael, C. (2001a). Modeling the interaction between soloist and accompaniment. In: *Proc. 14th Meeting of the FWO Research Society on Foundations of Music Research*, Ghent, Belgium.

Raphael, C. (2001b). Automated rhythm transcription. In: *Proc. International Symposium on Music Information Retrieval*, Indiana, 99–107.

Rosenthal, D.F. (1992). *Machine Rhythm: Computer emulation of human rhythm perception*. PhD thesis, Massachusetts Institute of Technology.

Rosenthal, D.F., & Okuno, H.G. (Eds.) (1998). *Computational auditory scene analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rossi, L. (1998). *Identification de sons polyphoniques de piano*. Ph.D. thesis, L'Université de Corse, Corsica, France.

Rowe, R. (2001). *Machine Musicianship*. Cambridge, MA: MIT Press.

Serra, X. (1997). Musical sound modeling with sinusoids plus noise. In: C. Roads, S. Pope, A. Picialli, & G. De Poli (Eds.), *Musical Signal Processing*. Lisse: Swets & Zeitlinger Publishers.

Sethares, W.A., & Staley, T.W. (1999). Periodicity transforms. *IEEE Trans. Signal Processing*, *47*, 2953–2964.

Sethares, W.A., & Staley, T.W. (2001). Meter and periodicity in musical performance. *Journal of New Music Research*, 22(5).

Sterian, A.D. (1999). Model-based segmentation of time-frequency images for musical transcription. Ph.D. thesis, University of Michigan.

Talkin, D. (1995). A robust algorithm for pitch tracking. In: W.B. Kleijn, & K.K. Paliwal (Eds.), *Speech Coding and Synthesis*. Amsterdam: Elsevier Science.

Temperley, D., & Sleator, D. (1999). Modeling meter and harmony: A preference-rule approach. *Computer Music Journal*, *23*, 10–27.

Temperley, D. (2001). *Cognition of Basic Musical Structures*. Cambridge, MA: MIT Press.

Tolonen, T., & Karjalainen, M. (2000). A computationally efficient multipitch analysis model. *IEEE Trans. Speech Audio Processing*, *8*, 708–716.

Viitaniemi, T., Klapuri, A., & Eronen, A. (2003). A probabilistic model for the transcription of single-voice melodies. In: *Proc. Finnish Signal Processing Symposium*, Tampere, Finland.

Virtanen, T. (2003). Sound source separation using sparse coding with temporal continuity objective. In: *Proc. International Computer Music Conference*, Singapore.

Wang, D.L., & Brown, G.J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. on Neural Networks*, 10(3).

Watson, C. (1985) *The computer analysis of polyphonic music*. Ph.D. thesis, University of Sydney.

Wu, M., Wang, D.L., & Brown, G.J. (2002). A multi-pitch tracking algorithm for noisy speech. In: *Proc. IEEE International Conference on Acoust., Speech, and Signal Processing*, Orlando, FL.

Zatorre, R.J., Belin, P., & Penhune, V.B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, *6*, 37–46.

Zils, A., Pachet, F., Delerue, O., & Gouyon, F. (2002). Automatic extraction of drum tracks from polyphonic music signals. In: *Proc. 2nd Int. Conference on Web Delivery of Music*, Darmstadt, Germany, 179–183.