J Intell Inf Syst
DOI 10.1007/s10844-013-0258-3

Automatic music transcription: challenges and future directions

Emmanouil Benetos · Simon Dixon · Dimitrios Giannoulis · Holger Kirchhoff · Anssi Klapuri

Received: 9 November 2012 / Revised: 26 June 2013 / Accepted: 27 June 2013

Abstract Automatic music transcription is considered by many to be a key enabling technology in music signal processing. However, the performance of transcription systems is still significantly below that of a human expert, and accuracies reported in recent years seem to have reached a limit, although the field is still very active. In this paper we analyse limitations of current methods and identify promising directions for future research. Current transcription methods use general purpose models which are unable to capture the rich diversity found in music signals. One way to overcome the limited performance of transcription systems is to tailor algorithms to specific use-cases. Semi-automatic approaches are another way of achieving a more reliable transcription. Also, the wealth of musical scores and corresponding

All authors contributed equally to this work.

E. Benetos (✉)
Department of Computer Science, City University London, London, UK
e-mail: emmanouil.benetos.1@city.ac.uk

S. Dixon · D. Giannoulis · H. Kirchhoff
Centre for Digital Music, Queen Mary University of London,
London, UK

S. Dixon
e-mail: simon.dixon@eecs.qmul.ac.uk

D. Giannoulis
e-mail: dimitrios.giannoulis@eecs.qmul.ac.uk

H. Kirchhoff
e-mail: holger.kirchhoff@eecs.qmul.ac.uk

A. Klapuri
Ovelin Ltd., Vilhonkatu 5, 00100 Helsinki, Finland

A. Klapuri
Tampere University of Technology, Korkeakoulunkatu 10,
33720 Tampere, Finland
e-mail: anssi.klapuri@tut.fi

Published online: 25 July 2013

audio data now available are a rich potential source of training data, via forced alignment of audio to scores, but large scale utilisation of such data has yet to be attempted. Other promising approaches include the integration of information from multiple algorithms and different musical aspects.

Keywords Music signal analysis · Music information retrieval · Automatic music transcription

1 Introduction

Automatic music transcription (AMT) is the process of converting an acoustic musical signal into some form of musical notation. In Cemgil (2004) it is defined as the process of converting an audio recording into a piano-roll notation (a two-dimensional representation of musical notes across time), while in Klapuri and Davy (2006) it is defined as the process of converting a recording into common music notation (i.e. a score). Even for expert musicians, transcribing polyphonic pieces of music is not a trivial task (see Chapter 1 of Klapuri and Davy 2006 and Klapuri et al. 2001), and while the problem of automatic pitch estimation for monophonic signals might be considered solved, the creation of an automated system able to transcribe polyphonic music without restrictions on the degree of polyphony or the instrument type still remains open.

The most immediate application of automatic music transcription is for allowing musicians to record the notes of an improvised performance in order to be able to reproduce it. AMT also has great value in musical styles where no score exists, e.g. music from oral traditions, jazz, pop, etc. In the past years, the problem of automatic music transcription has gained considerable research interest due to the numerous applications associated with the area, such as automatic search and annotation of musical information, interactive music systems (e.g. computer participation in live human performances, score following, and rhythm tracking), as well as musicological analysis (Bello 2003; Goto 2004; Klapuri and Davy 2006). An example of the transcription process can be seen in Fig. 1.

The AMT problem can be divided into several subtasks, which include: multi-pitch detection, note onset/offset detection, loudness estimation and quantisation, instrument recognition, extraction of rhythmic information, and time quantisation. The core problem in automatic transcription is the estimation of concurrent pitches in a time frame, also called multiple-F0 or multi-pitch detection.

In this work we address challenges and future directions for automatic transcription of polyphonic Western music, expanding upon the work presented in (Benetos et al. 2012). The related problem of melody transcription, i.e. the estimation of the predominant pitch, usually performed by a solo instrument or a lead singer, is not addressed in this paper; for an overview of melody transcription approaches the reader can refer to Poliner et al. (2007). Also, the field of content-based music information retrieval, which refers to automated processing of music for search and retrieval purposes and includes the AMT problem, is discussed in Casey et al. (2008). A recent state-of-the-art review of music signal analysis (which includes AMT) is given in Müller et al. (2011) while the work by Grosche et al. (2012) includes a recent state-of-the-art section on AMT systems.

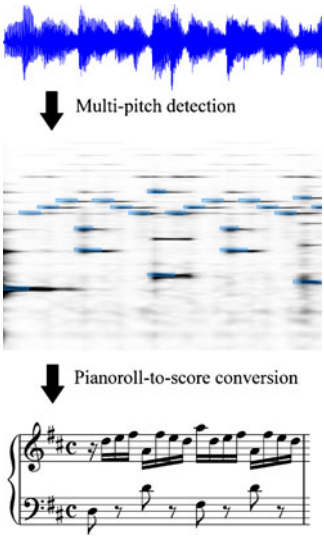

Fig. 1 An automatic music transcription example using the first bar of J.S. Bach's Prelude in D major. The *top panel* shows the time-domain audio signal, the *middle panel* shows a time-frequency representation with detected pitches superimposed, and the *bottom panel* shows the final score

2 State of the art

2.1 Multi-pitch detection and note tracking

In polyphonic music transcription, we are interested in detecting notes which might occur concurrently and could be produced by several instrument sources. The core problem for creating a system for polyphonic music transcription is thus multi-pitch estimation. The vast majority of AMT systems restrict their scope to performing multi-pitch detection and note tracking (either jointly or sequentially).

In Yeh (2008), multi-pitch detection systems were classified according to their estimation type as either joint or iterative. The iterative estimation approach extracts the most prominent pitch in each iteration, until no additional F0s can be estimated. Generally, iterative estimation models tend to accumulate errors at each iteration step, but are computationally inexpensive. On the contrary, joint estimation methods evaluate F0 combinations, leading to more accurate estimates but with increased computational cost. Recent developments in AMT show that the vast majority of proposed approaches now falls within the 'joint' category. Thus, the classification that will be presented in this paper organises multi-pitch detection systems according to the core techniques or models employed.

2.1.1 Feature-based multi-pitch detection

Most multiple-F0 estimation and note tracking systems employ methods derived from signal processing; a specific model is not employed, and notes are detected using audio features derived from the input time-frequency representation either in a joint or an iterative fashion. Typically, multiple-F0 estimation occurs using a pitch salience function (also called pitch strength function) or a pitch candidate set score function (Klapuri 2003; Pertusa and Iñesta 2008; Yeh 2008). These feature-based techniques have produced the best results in the Music Information Retrieval

Evaluation eXchange (MIREX) multi-F0 (frame-wise) and note tracking evaluations (Bay et al. 2009; Music Information Retrieval Evaluation eXchange 2011).

The best performing method in the MIREX multi-F0 and note tracking tasks for 2009–2011 was the work by Yeh (Yeh 2008), who proposed a joint pitch estimation algorithm based on a pitch candidate set score function. Given a set of pitch candidates, the overlapping partials are detected and smoothed according to the spectral smoothness principle, which states that the spectral envelope of a musical tone tends to be slowly varying as a function of frequency. The weighted score function for the pitch candidate set consists of 4 features: harmonicity, mean bandwidth, spectral centroid, and "synchronicity" (synchrony). A polyphony inference mechanism based on the score function increase selects the optimal pitch candidate set.

For 2012, the best performing method for the MIREX multi-F0 estimation and note tracking tasks was by Dressler (2012). As an input time/frequency representation, a multiresolution Fast Fourier Transform analysis is employed, where the magnitude for each spectral bin is multiplied with the bin's instantaneous frequency. Pitch estimation is made by identifying spectral peaks and performing pair-wise analysis on them, resulting on ranked peaks according to harmonicity, smoothness, the appearance of intermediate peaks, and harmonic number. Finally, the system tracks tones over time using an adaptive magnitude and a harmonic magnitude threshold.

Other notable feature-based AMT systems include the work by Pertusa and Iñesta (2008), who proposed a computationally inexpensive method for multi-pitch detection which computes a pitch salience function and evaluates combinations of pitch candidates using a measure of distance between a harmonic partial sequence (HPS) and a smoothed HPS. Another approach for feature-based AMT was proposed in Reis et al. (2008), which uses genetic algorithms for estimating a transcription by mutating the solution until it matches a similarity criterion between the original signal and the synthesized transcribed signal. More recently, Grosche et al. (2012) proposed an AMT method based on a mid-level representation derived from a multiresolution Fourier transform combined with an instantaneous frequency estimation. The system also combines onset detection and tuning estimation for computing frame-based estimates. Finally, Nam et al. (2011) proposed a classification-based approach for piano transcription using features learned from deep belief networks (Humphrey et al. 2013) for computing a mid-level time-pitch representation.

2.1.2 Statistical model-based multi-pitch detection

Many approaches in the literature formulate the multiple-F0 estimation problem within a statistical framework. Given an observed frame $\mathbf{x}$ and a set $\boldsymbol{\mathcal{C}}$ of all possible fundamental frequency combinations, the frame-based multiple-F0 estimation problem can then be viewed as a maximum a posteriori (MAP) estimation problem (Emiya et al. 2010):

$$\hat{\mathcal{C}}_{MAP} = \arg\max_{\mathcal{C}\in\boldsymbol{\mathcal{C}}} P(\mathcal{C}|\mathbf{x}) = \arg\max_{\mathcal{C}\in\boldsymbol{\mathcal{C}}} \frac{P(\mathbf{x}|\mathcal{C})\,P(\mathcal{C})}{P(\mathbf{x})} \tag{1}$$

where $\mathcal{C} = \{F_0^1, \ldots, F_0^N\}$ is a set of fundamental frequencies, $\boldsymbol{\mathcal{C}}$ is the set of all possible F0 combinations, and $\mathbf{x}$ is the observed audio signal within a single analysis frame.

An example of MAP estimation-based transcription is the *PreFEst* system (Goto 2004), where each harmonic is modelled by a Gaussian centered at its position on the log-frequency axis. MAP estimation is performed using the expectation-maximisation (EM) algorithm. An extension of the method from Goto (2004) was proposed by Kameoka et al. (2007), called harmonic temporal structured clustering (HTC), which jointly estimates multiple fundamental frequencies, onsets, offsets, and dynamics. Partials are modelled using Gaussians placed at the positions of partials in the log-frequency domain and the synchronous evolution of partials belonging to the same source is modelled by Gaussian mixtures.

If no prior information is specified, the problem can be expressed as a maximum likelihood (ML) estimation problem using Bayes' rule (e.g. Cemgil et al. 2006, Emiya et al. 2010):

$$\hat{\mathcal{C}}_{ML} = \arg\max_{\mathcal{C}\in\boldsymbol{\mathcal{C}}} P(\mathbf{x}|\mathcal{C}) \tag{2}$$

It should be noted that the MAP estimator of (1) is equivalent to the ML estimator of (2) if no prior information on the F0 mixtures is specified.

A time-domain Bayesian approach for AMT which used a Gabor atomic model was proposed in Davy et al. (2006), which used a Markov chain Monte Carlo (MCMC) method for inference, while the model also supported time-varying amplitudes and inharmonicity. An ML approach for multi-pitch detection which models spectral peaks and non-peak regions was proposed by Duan et al. (2010). The likelihood function of the model is composed of the peak region likelihood (probability that a peak is detected in the spectrum given a pitch) and the non-peak region likelihood (probability of not detecting any partials in a non-peak region), which are complementary. Emiya et al. (2010) proposed a joint estimation method for piano notes using a likelihood function which models the spectral envelope of overtones using a smooth autoregressive model and models the residual noise using a low-order moving average model.

More recently, Peeling and Godsill (2011) also proposed a likelihood function for multiple-F0 estimation where for a given time frame, the occurrence of peaks in the frequency domain is assumed to follow an inhomogeneous Poisson process. Also, Koretz and Tabrikian (2011) proposed an iterative method for multi-pitch estimation, which combines MAP and ML criteria. The predominant source is expressed using a harmonic model while the remaining harmonic signals are modelled as Gaussian interference sources. Finally, a nonparametric Bayesian approach for AMT was proposed in Yoshii and Goto (2012), where a statistical method called Infinite Latent Harmonic Allocation (iLHA) was proposed for detecting multiple fundamental frequencies in polyphonic audio signals, eliminating the problem of fixing the number of parameters.

2.1.3 Spectrogram factorisation-based multi-pitch detection

The majority of recent multi-pitch detection papers utilise and expand *spectrogram factorisation* techniques. Non-negative matrix factorisation (NMF) is a technique first introduced as a tool for music transcription in Smaragdis and Brown (2003).

In its simplest form, the NMF model decomposes an input spectrogram $\mathbf{X} \in \mathbb{R}_{+}^{K \times N}$ with K frequency bins and N frames as:

$$\mathbf{X} \approx \mathbf{WH} \tag{3}$$

where $R << K, N$; $\mathbf{W} \in \mathbb{R}_{+}^{K \times R}$ contains the spectral bases for each of the R pitch components; and $\mathbf{H} \in \mathbb{R}_{+}^{R \times N}$ is the pitch activity matrix across time.

Applications of NMF for AMT include the work by Cont (2006), where sparseness constraints were added into the NMF update rules, in an effort to find meaningful transcriptions using a minimum number of non-zero elements in $\mathbf{H}$. Vincent et al. (2010) incorporated harmonicity constraints in the NMF model, resulting in two algorithms: harmonic and inharmonic NMF. The model additionally constrains each basis spectrum to be expressed as a weighted sum of narrowband spectra, in order to preserve a smooth spectral envelope for the resulting basis functions. The inharmonic version of the algorithm is also able to support deviations from perfect harmonicity and standard tuning. Also, Bertin et al. (2010) proposed a Bayesian framework for NMF, which considers each pitch as a model of Gaussian components in harmonic positions. Spectral smoothness constraints are incorporated into the likelihood function, and for parameter estimation the space alternating generalised EM algorithm (SAGE) is employed. More recently, Ochiai et al. (2012) proposed an algorithm for multi-pitch detection and beat structure analysis. The NMF objective function is constrained using information from the rhythmic structure of the recording, which helps improve transcription accuracy in highly repetitive recordings.

An alternative formulation of NMF called probabilistic latent component analysis (PLCA) has also been employed for transcription. In PLCA (Smaragdis et al. 2006) the input spectrogram is considered to be a bivariate probability distribution which is decomposed into a product of one-dimensional marginal distributions. An extension of the PLCA algorithm was used for multiple-instrument transcription in Grindlay and Ellis (2011), where a system was proposed which supported multiple spectral templates for each pitch and instrument source. The notion of *eigeninstruments* was used for modelling fixed spectral templates as a linear combination of basic instrument models. A model that extended the convolutive PLCA algorithm was proposed in Benetos and Dixon (2012), which incorporated shifting across log-frequency for supporting frequency modulations, as well as the use of multiple spectral templates per pitch and per instrument source. Also, Fuentes et al. (2011) extended the convolutive PLCA algorithm, by modelling each note as a weighted sum of narrowband log-spectra which are also shifted across log-frequency.

Sparse coding techniques employ a linear model similar to the NMF model of (3), but instead of assuming non-negativity, it is assumed that the sources are non-active most of the time, resulting in a sparse matrix $\mathbf{H}$. In order to derive the bases, ML estimation is performed. Abdallah and Plumbley (2004) used an ML approach for dictionary learning using non-negative sparse coding. Dictionary learning occurs directly from polyphonic samples, without requiring training on monophonic data. Bertin et al. (2007) employed the non-negative k-means singular value decomposition algorithm (NKSVD) algorithm for multi-pitch detection, comparing its performance with the NMF algorithm. More recently in O'Hanlon et al. (2012), structured sparsity (also called group sparsity) was applied to piano transcription. In group sparsity, groups of atoms tend to be active at the same time. Also, sparse

coding of Fourier coefficients was used in Lee et al. (2012), which solves the sparse representation problem using l_1 minimisation and utilises exemplars for training.

2.1.4 Note tracking

Typically AMT algorithms compute a time-pitch representation which needs to be further processed in order to detect note events with a discrete pitch value, an onset time and an offset time. This procedure is called *note tracking* or *note smoothing*. Most spectrogram factorisation-based methods estimate the binary piano-roll representation from the pitch activation matrix using simple thresholding (Grindlay and Ellis 2011; Vincent et al. 2010).

One simple and fast solution for note tracking is minimum duration pruning (Dessein et al. 2010), which is applied after thresholding. Essentially, note events which have a duration smaller than a predefined value are removed from the final piano-roll. This method was also used in Bello et al. (2006), where more complex rules for note tracking were used, addressing cases such as where a small gap exists between two note events.

Hidden Markov models (HMMs) are frequently used at a postprocessing stage for note tracking. In Poliner and Ellis (2007), a note tracking method was proposed using pitch-wise HMMs, where each HMM has two states, denoting note activity and inactivity. The HMM parameters (state transitions and priors) were learned directly from a ground-truth training set, while the observation probability is given by the posteriogram output for a specific pitch. In Ryynänen and Klapuri (2005) a feature-based multi-pitch detection system was combined with a musicological model for estimating musical key and note transition probabilities. Note events are described using 3-state HMMs, which model the attack, sustain, and noise/silence states of each sound. Information from an onset detection function was also incorporated. In addition, context-dependent HMMs were employed in Grosche et al. (2012) for determining note events by combining the output of a multi-pitch detection system with an onset detection system.

Finally, dynamic Bayesian networks (DBNs) were proposed in Raczyński et al. (2009) for note tracking using as input the pitch activation of an NMF-based multi-pitch detection algorithm. The DBN has a note layer in the lowest level, followed by a note combination layer. Model parameters were learned using MIDI files from F. Chopin piano pieces.

2.2 Other transcription subtasks

For an AMT system to output complete music notation, it has to solve a set of problems, central to which is multi-pitch estimation (see Section 2.1). The other subtasks involve the estimation of features relating to rhythm, melody, harmony and instrumentation, which carry information which, if integrated, could improve transcription performance. For many of these descriptors, their estimation has been studied in isolation, and we briefly review some of the most relevant contributions to instrument recognition, detection of onsets and offsets, extraction of rhythmic information (tempo, beat, and musical timing), and estimation of pitch and harmony (key, chords and pitch spelling).

Instrument recognition or identification attempts to identify the musical instrument(s) playing in a music excerpt or piece. Early work on the task involved monophonic musical instrument identification, where only one instrument was playing at a given time Herrera-Boyer et al. (2006). In most music, however, instruments do not play in isolation and therefore multiple-instrument (or polyphonic) identification is necessary. Instrument identification in a polyphonic context is rendered difficult by the way the different sources blend with each other, resulting in a high degree of overlap in the time-frequency domain. The task is closely related to sound source separation and as a result, many systems operate by first separating the signals of different instruments from the mixture and then classifying them separately (Bay and Beauchamp 2012; Burred et al. 2009; Heittola et al. 2009). The benefit of this approach is that the classification is performed on isolated instruments, thus is likely to have better results, assuming that the demanding source separation step is successful.

There are also systems that try to extract features directly from the mixture. In (Little and Pardo 2008), the authors used weakly-labelled audio mixtures to train binary classifiers for instrument detection, whereas in Barbedo and Tzanetakis (2011), the proposed algorithm extracted features by focusing on time-frequency regions with isolated note partials. In Kitahara et al. (2007), the authors introduced a note-estimation-free instrument recognition system that made use of a spectrogram-like representation (Instrogram). A series of approaches incorporate missing feature theory and aim to generate time-frequency masks that indicate spectrotemporal regions that belong only to a particular instrument which can then be classified more accurately since regions that are corrupted by noise or interference are kept out of the classification process (Eggink and Brown 2003; Giannoulis and Klapuri 2013). Lastly, a third category includes systems that try to jointly separate and recognise the instruments of the mixture by employing parametric signal models and probabilistic inference (Itoyama et al. 2011; Wu et al. 2011) or by utilizing a mid-level representation of the signal and trying to model it as a sum of instrument- and pitch-specific active atoms (Bay and Beauchamp 2012; Leveau et al. 2008).

Onset detection (finding the beginnings of notes or events) is the first step towards understanding the underlying periodicities and accents in the music, which ultimately define the rhythm. Although most transcription systems do not yet attempt to interpret the timing of notes with respect to an underlying metrical structure, onset detection has a large impact on transcription results, due to the way note tracking is usually evaluated. There is no unique way to characterise onsets, but some common features of onsets can be listed, such as a sudden burst of energy or change of harmonic content in the signal, or unpredictable and unstable components followed by a steady-state region. Onsets are difficult to identify directly from time-domain signals, particularly in polyphonic and multi-instrumental musical signals, so it is usual to compute an intermediate representation, called an *onset detection function*, which quantifies the amount of change in the signal properties from frame to frame. Onset detection functions are typically computed from frequency-domain signals, using the band-wise magnitude and/or phase to compute spectral flux, phase deviation or complex domain detection functions (Bello et al. 2005; Dixon et al. 2006). Onsets are then computed from the detection function by peak-picking with suitable thresholds and constraints. Other onset detection methods that have performed well in MIREX evaluations include the use of psychoacoustically motivated features (Collins 2005),

transient peak classification (Röbel 2005) and pitch-based features (Zhou and Reiss 2007). A data-driven approach using supervised learning, where various neural network architectures have been utilised, has given the best results in several MIREX evaluations, including the most recent one (2012) (Böck et al. 2012; Eyben et al. 2012; Lacoste and Eck 2007). Finally, Degara et al. (2011) exploit rhythmic regularity in music using a probabilistic framework to improve onset detection, showing that the integration of onset detection with higher-level rhythmic processing is advantageous.

Considerably less attention has been given to the detection of offsets, or ends of notes. The task itself is ill-defined, particularly for percussive instruments, where the partials decay exponentially and it is not possible to state unambiguously where a note ends, especially in a polyphonic context. Offset detection is also less important for rhythmic analysis, since the tempo and beat structure can be determined from onset times without reference to any offsets. So it is mainly in the context of transcription that offset detection has been considered. For threshold-based approaches, the offset is usually defined by a threshold relative to the maximum level of the note. Other approaches train a hidden Markov model with two states (on and off) to detect both offsets for each pitch Benetos and Dixon (2011).

The temporal organisation of most Western music is centred around a metrical structure consisting of a hierarchical set of pulses, where a pulse is a regularly spaced sequence of accents (or *beats*) in time. In order to interpret an audio recording in terms of such a structure (which is necessary in order to produce Western music notation), the first step is to determine the rate of the most salient pulse (or some measure of its central tendency), which is called the *tempo*. Algorithms used for tempo induction include autocorrelation, comb filterbanks, inter-onset interval histograms, Fourier transforms, and periodicity transform, which are applied to audio features such as an onset detection function (Gouyon and Dixon 2005). The next step involves estimating the timing of the beats constituting the main pulse, a task known as *beat tracking*. Again, numerous approaches have been proposed, such as rule-based methods (Desain and Honing 1999), adaptive oscillators (Large and Kolen 1994), agent-based or multiple hypothesis trackers (Dixon 2001), filter-banks (Davies and Plumbley 2007), dynamical systems (Cemgil and Kappen 2003) and probabilistic models (Degara et al. 2012). Beat tracking methods are evaluated in Gouyon et al. (2006), McKinney et al. (2007). The final step for metrical analysis consists of inferring the time signature, which indicates how beats are grouped and subdivided at respectively higher and lower metrical levels, and assigning (quantising) each onset and offset time to a position in this metrical structure (Cemgil and Kappen 2003).

Most Western music also has a harmonic organisation around a tonal centre and scale (or mode), which together define the *key* of the music. The key is generally stable over whole, or at least sections of, musical pieces. At a local level, the harmony is described by *chords*, which are combinations of simultaneous, sequential or implied notes which are perceived to belong together and have more than a transitory function. Algorithms for key detection use template matching (Izmirli 2005) or hidden Markov models (HMMs) (Noland and Sandler 2006; Peeters 2006), and the audio is converted to a mid-level representation such as chroma or pitch class vectors. Chord estimation methods similarly use template matching (Oudre et al. 2009) and HMMs (Lee and Slaney 2008), and several approaches jointly estimate other variables such as key, metre and bassline (Mauch and Dixon 2010; Papadopoulos and Peeters 2008; Ryynänen and Klapuri 2008) in a probabilistic framework such as a dynamic Bayesian network.

3 Challenges

Despite significant progress in AMT research, there exists no end-user application that can accurately and reliably transcribe music containing the range of instrument combinations and genres found in recorded music. The performance of even the most recent systems is still clearly below that of a human expert, despite the fact that humans themselves produce imperfect results and require multiple takes, while making extensive use of prior knowledge and complex inference. Furthermore, current test sets are limited in their complexity and coverage. Table 1 gives the results for the frame-based multiple-F0 estimation task of the MIREX evaluation (Music Information Retrieval Evaluation eXchange 2011). These highlight the stagnation in performance of which we speak. It is also worth mentioning that the best algorithm proposed by Yeh (2008) (who also provided a subset of the test dataset) has gone unimproved since 2009.

Results for the note tracking task over the years are presented in Table 2. These are much inferior, especially for the case when both onset and offset detection is taken into account for the computation of the metrics. A notable exception among them is the algorithm proposed by Dressler (2012) which performs exceptionally well for the task with F-measures of 0.45 and 0.65, respectively for the two note tracking tasks, bringing the system's performance up to the levels attained for multiple-F0 estimation, but not higher. A possible explanation behind the improved performance of the algorithm could be the more sophisticated note tracking algorithm that is based upon perceptual studies, whereas the standard note tracking systems are simply filtering the note activations.

The observed plateau in AMT system performance can be further emphasized when we compare multiple-instrument transcription with piano transcription. The results for the best systems on the note tracking task (with onset only detection) fluctuate around 0.60 over the years with Dressler's algorithm obtaining the best result, measured at 0.66 in the 2012 evaluation which is almost equivalent to that for the multiple instrument transcription task. It should however be noted that the dataset used for the piano note tracking task consists of real polyphonic piano recordings generated using a disklavier playback piano and not artificially synthesized pieces using RWC MIDI and RWC musical instrument samples to create the polyphonic mixtures used for the multiple-instrument transcription note tracking task Music Information Retrieval Evaluation eXchange (2011).

The shortcomings of existing methodologies do not stop here. Currently proposed systems also fall short in flexibility to deal with diverse target data. Music genres

Table 1 Best results using the accuracy metric for the MIREX Multi-F0 estimation task, from 2009–2012

Participants	2009	2010	2011	2012
Yeh and Röbel	0.69	0.69	0.68	–
Dressler	–	–	0.63	0.64
Benetos and Dixon	–	0.47	0.57	0.58
Duan et al.	0.57	0.55	–	–
Fuentes et al.	–	–	–	0.56

Details about the employed metric can be found in Music Information Retrieval Evaluation eXchange (2011)

Table 2 Best results usin the avg. F-measure (onset detection only and onset-offset detection respectively) for the MIREX Multi-F0 note tracking task, from 2009–2012

Participants	2009	2010	2011	2012
Avg. F-measure (Onset only)				
Yeh and Röbel	0.50	0.53	0.56	–
Dressler	–	–	–	0.65
Benetos and Dixon	–	–	0.45	0.43
Duan et al.	0.43	0.41	–	–
Fuentes et al.	–	–	–	0.61
Avg. F-measure (Onset-Offset)				
Yeh and Röbel	0.31	0.33	0.35	–
Dressler	–	–	–	0.45
Benetos and Dixon	–	–	0.21	0.23
Duan et al.	0.22	0.19	–	–
Fuentes et al.	–	–	–	0.39

Details about the employed metric can be found in Music Information Retrieval Evaluation eXchange (2011)

like classical, hip-hop, ambient electronic and traditional Chinese music have little in common. Furthermore styles of notation vary with genre. For example Pop/Rock notation might represent melody, chords and (perhaps) bass line, whereas a classical score would usually contain all the notes to be played, and electroacoustic music has no standard means of notation. Similarly, the parts for specific instruments might require additional notation details like playing style (e.g. pizzicato) and fingering. The user's expectations of a transcription system depend on notational conventions specific to the instrument and style being transcribed. The task of tailoring AMT systems to specific styles has yet to be addressed in the literature.

Typically, algorithms are developed independently to carry out individual tasks such as multiple-F0 detection, beat tracking and instrument recognition. Although this is necessary, considering the complexity of each task, the challenge remains to combine the outputs of the algorithms, or better, the algorithms themselves, to perform joint estimation of all parameters, in order to avoid the cascading of errors when algorithms are combined sequentially.

Another challenge concerns the availability of data for training and evaluation. Although there is no shortage of transcriptions and scores in standard music notation, human effort is required to digitise and time-align them to recordings. Except for the case of solo piano where available data include the MAPS database (Emiya et al. 2010) and the Disklavier piano dataset (Poliner and Ellis 2007), although the latter is synthesized from MIDI files extracted from the Disklavier performance, data sets currently employed for evaluation are small: a subset of the RWC database (Goto et al. 2002) which contains only twelve 30-s segments is commonly used (although the RWC database contains many more recordings) and the MIREX multi-F0 development set lasts only 54 s. Such small datasets cannot be considered representative; the danger of overfitting and thus overestimating system performance is high. It has been observed for several tasks that dataset developers tend to attain the best MIREX results (Music Information Retrieval Evaluation eXchange 2011).

At present, no single unifying framework has been established for music transcription in the way that HMMs have been for speech recognition. Instead, there

are multiple approaches. Among them, spectrogram factorisation is rapidly growing in popularity and could potentially establish itself as the mainstream, even though at present a large number of approaches involve the use of signal processing and feature extraction based techniques. Spectrogram factorisation techniques are mainly frame-based even though they can take into account temporal evolution of notes and global signal statistics. Other approaches that would treat notes as time-frequency objects and exploit dynamic time warping or HMMs integrated at a low level could offer a breath of fresh air on research in the field. Likewise, there is no standard method for front end processing of the signal, with various approaches including the short-time Fourier transform, constant-Q transform (Brown 1991) and auditory models, each leading to different mid-level representations. The challenge in this case is to characterise the impact of such design decisions on AMT results.

In addition to the above, the research community shares code and data on an ad hoc basis, which limits or forbids entirely the level of re-use of research outputs. The lack of standard methodology is also a contributing factor, making it difficult to develop a useful shared code-base. The Reproducible Research movement (Buckheit and Donoho 1995), with its emphasis on open software and data, provides examples of best practice which are worthy of consideration by the MIR community. Vandewalle et al. (2009) cite the benefits to the scientific community when research is performed with reproducibility in mind, and well documented code and data are made publicly available: it facilitates building upon others' work, and allows researchers to spend more time on novel research rather than reimplementing existing ideas, algorithms and code. To support this, they present evidence showing that highly cited papers typically have code and data available online. Other than that, it is very hard to perform a direct and objective comparison between open-source software or algorithms and a proprietary equivalent. From limited comparative experiments one can find in the literature, it is not possible to claim which exhibits higher quality or "better software" Oram and Wilson (2010)(Ch.15). However, we can argue that writing open-source code promotes some aspects of what are "good programming practices" Wilson et al. (2012), while also promoting the inclusion of more extensive and complete documentation, modularization, and version control that are shown to improve the productivity of scientific programming (Oram and Wilson 2010; Wilson et al. 2012).

Finally, present research in AMT introduces certain challenges in itself that might constrain the evolution of the field. Advances in AMT research have mainly come from engineers and computer scientists, particularly those specialising in machine learning. Currently there is minimal contribution from computational musicologists, music psychologists or acousticians. Here the challenge is to integrate knowledge from these fields, either from the literature or by engaging these experts as collaborators in AMT research and creating a stronger bond between the MIR community and other fields.

AMT research is quite active and vibrant at present, and we do not presume to predict what the state of the art will be in the next years and decades. In the remainder of the paper we propose promising techniques that could be utilised and further investigated, with some of them having been so already, in order to address the aforementioned limitations in transcription performance. Figure 2 depicts a general architecture of a transcription system, incorporating techniques discussed in the following sections. In the core of the system lie the multi-pitch detection and note

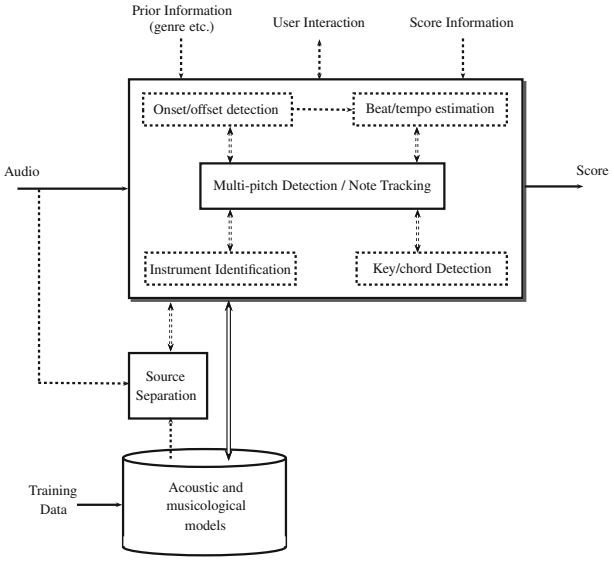

Fig. 2 Proposed general architecture of a music transcription system. Optional subsystems and algorithms are presented using *dashed lines*. The *double arrows* highlight connections between systems that include fusion of information and a more interactive communication among the systems

tracking algorithms. Four transcription sub-tasks related to multi-pitch detection and note tracking appear as optional system algorithms (dotted boxes) that can be integrated into a transcription system. These are: instrument identification, key and chord estimation, onset and offset detection, and tempo and beat estimation. Source separation, an independent but interrelated problem, could be addressed with a separate system that could inform and interact with the transcription system in general, and more specifically with the instrument identification subsystem. Optionally, information can also be fed externally to the transcription system. This could be given as prior information (i.e. genre, instrumentation, etc.), via user-interaction or by providing information from a partially correct or incomplete pre-existing score. Finally, training data can be utilized to learn acoustic and musicological models which subsequently inform and interact with the transcription system.

4 Informed transcription

4.1 Semi-automatic approaches

The fact that current state-of-the-art AMT systems do not reach the same level of accuracy as transcriptions made by human experts gives rise to the question of whether, and how, a human user could assist the computational transcription process in order to attain satisfactory transcription results. Certain skills possessed by human listeners, such as instrument identification, note onset detection and auditory stream segregation, are crucial for an accurate transcription of the musical content, but are often difficult to model algorithmically. Computers, on the other hand, are capable of performing tasks quickly, repeatedly and on large amounts of data. Combining human knowledge and perception with algorithmic approaches could thus lead to

transcription results that are more accurate than fully-automatic transcriptions and that are obtained in a shorter time than a human transcription. We refer to these approaches as *semi-automatic* or *user-assisted transcription* systems. Involving the user in the transcription process entails that these systems are not applicable to the analysis of large music databases. Such systems can, however, be useful when a more detailed and accurate transcription of individual music pieces is required and potential users could hence be musicologists, arrangers, composers and performing musicians.

The main challenges of user-assisted transcription systems are to identify areas in which human input can be beneficial for the transcription process, and to integrate the high-level human knowledge into the low-level signal analysis. Different types of user information might thereby require different ways of incorporating that knowledge, which might include the application of user feedback loops in order to refine the estimation of individual low-level parameter estimates. Further challenges include the more practical aspects such as interface design and minimising the amount and complexity of information required of users.

Criteria for the user input include the fact that the input needs to provide information that otherwise could not be easily inferred algorithmically. Any required input also needs to be reliably extractable by the user, who might not be an expert musician, and it should not require too much time and effort from the user to provide that information. In principle any acoustic or score-related information that matches the criteria above can act as prior information for the system. Depending on the expertise of the targeted users, this information could include key, tempo and time signature of the piece, structural information, information about the instrument types in the recording, or even asking the user to label a few chords or notes for each instrument. Although many proposed transcription systems—often silently—make assumptions about certain parameters, such as the number or types of instruments in the recording (e.g. Dessein et al. 2010, Grindlay and Ellis 2011, Lee et al. 2012), not many systems explicitly incorporate prior information from a human user.

As an example, in Kirchhoff et al. (2012), two different types of user information were compared in a user-assisted music transcription system: naming the instrument types in the recording, and labelling notes for each instrument. In the first case, previously learnt spectra of the same instrument types were used for the decomposition of the time-frequency representation, whereas in the second case, instrument spectra were derived directly from the instruments in the recording under analysis based on the user labels. The results (cf. Fig. 3) showed considerably better accuracies for the second case, across the full range of numbers of instruments in the target mixture. Similarly, Fuentes et al. (2012) asked the user to highlight notes in a mid-level representation in order to separate the main melody. Smaragdis and Mysore (2009) enabled the user to specify the melody to extract by humming along to the music. This knowledge enabled the authors to sidestep the error-prone tasks of source identification and timbre modelling. A transcription system that post-processes the transcription result based on user-input was proposed by Dittmar and Abeßer (2008). It allowed users to automatically snap detected notes to a detected beat grid and to the diatonic scale of the user-specified key. This feature of the system was not evaluated.

Finally, other tasks (or fields of research) have incorporated user-provided prior information as a method to improve overall performance. In the context of source

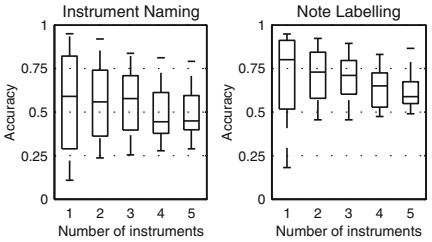

Fig. 3 Achieved accuracies of a user-assisted transcription system as a function of the number of instruments in the mixture. The *left panel* shows results for the case where instrument types were provided by the user. In the *right panel*, the user labelled notes for each instrument

separation, Ozerov et al. (2012) proposed a framework that enables the incorporation of prior knowledge about the number and types of sources, and the mixing model. The authors showed that by using prior information, a better separation could be achieved than with a completely blind system. A future challenge could be the development of a similar framework for incorporating prior information for user-assisted transcription.

In addition to their practical use as interactive systems, user-assisted transcription systems might also pave the way for more robust fully-automatic systems, because they allow algorithms to focus on a subset of the required tasks while at the same time being able to revert to reliable information from other subtasks (cf. Section 6). This enables isolated evaluation of the proposed solutions in an integrated framework.

4.2 Score-informed approaches

Contrary to speech, only a fraction of Western music is fully spontaneous, as musical performances are typically based on an underlying composition or song. Although transcription is usually associated with the analysis of an unknown piece, there are certain applications for which a score is available, and in these cases the AMT system can exploit this additional knowledge (Scheirer 1997) in order to help us understand the relationship between score and audio. This *score-informed transcription* area has certain similarities to the emerging topic of informed source separation (see also Section 6.3).

One application area where a score is available is automatic instrument tutoring (Benetos et al. 2012; Dittmar et al. 2012; Wang and Zhang 2008), where a system evaluates the performance of a student based on a reference score and provides feedback. Thus, the correctly played passages need to be identified, along with any mistakes made by the student, such as missed or extra played notes. An example of a score-informed transcription for automatic piano tutoring is given in Fig. 4. In Benetos et al. (2012) it was shown that the score-informed system was able to detect correct and extra notes played by students, but had a considerably lower performance regarding missing notes. Another challenge for score-informed transcription is how to treat structural errors in a piece, i.e. major changes in a performance and not local mistakes. This would require a robust alignment algorithm operating within the score-informed transcription framework.

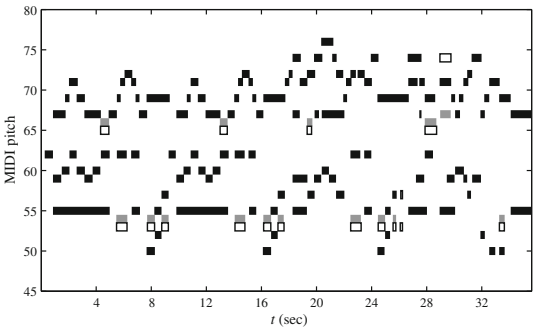

Fig. 4 The score-informed piano transcription of a performance of J. Brahms' *The Sandman*, from Benetos et al. (2012). *Black* corresponds to correct notes, *gray* to missed notes and empty rectangles to extra notes played by the student

Another example application is the analysis of expressive performance, where the tempo, dynamics, articulation and timing relative to the score are the focus of the analysis. There are often small differences between the reference score and the performance (e.g. ornamentation), and in most cases, the score will not contain the absolute timing of notes and thus will need to be time-aligned with the recording as a first step.

One way to utilise the automatically-aligned score is for initialising the pitch activity matrix **H** in a spectrogram factorisation-based model (see (3)), and keeping these fixed while the spectral templates **W** are learned, as in Ewert and Müller (2012). After the templates are learned, the gain matrix could also be updated in order to cater for note differences between the score and the recording.

5 Instrument- and genre-specific transcription

Current AMT approaches usually employ instrument models that are not restricted to specific instrument types, but are applicable and adaptable to a wide range of musical instruments. In fact, most transcription algorithms that are based on heuristic rules and those that employ perceptual models even deliberately disregard specific timbral characteristics in order to enable an instrument-independent detection of notes. Even many transcription methods that aim to transcribe solo piano music are not so much tailored to piano music as *tested* on such music; these approaches do not necessarily implement a piano-specific instrument model. Similarly, the aim of many transcription methods is to be applicable to a broad range of musical genres.

The fact that only a small number of publications on instrument- and genre-specific transcription exist, is particularly surprising when we compare AMT to the more mature discipline of automatic speech recognition. Continuous speech recognition systems are practically always language-specific and typically also domain-specific, and many modern speech recognisers include speaker adaptation (Huang et al. 2001).

Transcription systems usually try to model a wide range of musical instruments using a single set of computational methods, thereby assuming that those methods can be applied equally well to different kinds of instruments. A prominent example is the non-negative matrix factorisation technique (cf. Section 2.1.3) which can be used to find prototype spectra for the different pitches in the recording that capture the instrument-specific average harmonic partial amplitudes (e.g. Dessein et al.

2010). However, depending on the sound production mechanism of instruments, their characteristics can differ considerably and might not be captured equally well by the same computational model or might at least require defining a set of instrument-specific parameters and constraints in the common model used. The NMF technique for example would require additional computational complexity and time by introducing more than a single basis element per pitch per instrument in order to account for any variations in the partial amplitudes during the course of a note or due to differences in dynamic levels which might have a considerable effect on the transcription accuracy.

Furthermore, acoustic instruments incorporate a wide range of playing styles, which can differ notably in sound quality. To model these differences we can turn to the extensive literature on the physical modelling of musical instruments. A promising direction could be to incorporate these models in the transcription process and adapt their specific parameters to the recording under analysis. Some examples of instrument-specific transcription can be found for violin (Barbancho et al. 2009; Loscos et al. 2006), bells (Marolt 2012), tabla (Gillet and Richard 2003) and guitar (Barbancho et al. 2012). The application of instrument-specific models, however, requires the target instrumentation either to be known or inferred from the recording via instrument recognition algorithms (cf. Section 2.2).

Recently, the increasing interest of the MIR community in the application of music analysis techniques to non-Western music has underlined the fact that different musical genres require different analysis techniques in order to be able to extract genre-specific musical structures (e.g. Özaslan et al. 2012). Restricting a transcription system to a certain musical genre enables the incorporation of specific (expert) knowledge about that genre. Musicological knowledge about structure (e.g. sonata form), harmony progressions (e.g. 12-bar blues) or specific instruments could for example be used to enhance transcription accuracy. Genre-specific AMT systems have been designed for genres such as Australian aborginal music (Nesbit et al. 2004), but genre-specific methods could likewise be applied to other Western and non-Western musical genres. In order to build a general-purpose AMT system, several genre-specific transcription systems could be combined and selected based on a preliminary genre classification stage.

6 Information integration

6.1 Fusing information across the aspects of music

Many systems for note tracking combine multiple-F0 estimation with onset and offset detection, but disregard concurrent research on other aspects of music, for example the estimation of various music content descriptors such as instrumentation, rhythm, or tonality. These descriptors are highly interdependent and they could be analysed jointly, combining information across time and across features to improve transcription performance. This, for example, can be seen clearly from the latest MIREX evaluation results (Music Information Retrieval Evaluation eXchange 2011), where independent estimators for various musical aspects apart from onset detection, such as, key detection and tempo estimation have performances around 80 % and could potentially improve the transcription process if integrated in an AMT system.

A human transcriber interprets the performed notes in the context of the metrical structure. Extensive research has been performed into beat tracking and rhythm parsing (Gouyon and Dixon 2005), but transcription rarely takes advantage of this knowledge. An exception is the transcription system in Kameoka et al. (2012), which combines the processes of tempo and note onset estimation with that of note extraction, using a generative model for polyphonic spectrograms that simultaneously extracts the notes and the structure of a piece based on a 2-dimensional hierarchical tree-structured Bayesian model. However the system's performance has not been systematically evaluated and there are some computational complexity issues with this approach. Another case is Ochiai et al. (2012), where the proposed system explicitly models the beat structure information and extracts a set of musically meaningful temporal constraints that are subsequently applied on an NMF-based transcription model in order to improve the overall transcription performance. The system greatly improves over the unconstrained NMF but would have to be tested more rigorously to observe its full potential.

Another example, this time on chord transcription with the use of beat-synchronous features, is Mauch and Dixon (2010), where the audio is segmented according to the location of beats, and features are averaged over these beat-length intervals. The advantage of a more robust feature (less overlap between succeeding chords) is balanced by a loss in temporal resolution (harmonic change is assumed not to occur within a beat). For note transcription, it is unrealistic to assume that notes do not change within beats, but a promising approach would be to use a similar technique at a lower (i.e. sub-beat) metrical level, corresponding to the fastest note sequences. The resulting features would be more robust than frame-level features, and advantage could be taken of known (or learnt) rhythmic patterns and effects of metrical position.

Key is another high-level musical cue that, if known or estimated, provides useful prior information for the extraction of notes and chords. As described in Section 2.2, key can be modelled as imposing a probability distribution over notes and chords for different metrical positions and durations. Therefore, by specifically modelling key, transcription accuracy can be improved, e.g. by giving more weight to notes which belong to the current key. Genre and style are also influential factors for modelling the distribution of pitch classes in a key. The key estimation approaches that have been proposed are rarely exploited for AMT, with one exception being that of Ryynänen and Klapuri (2005), which gave the best results for the MIREX 2008 note tracking task.

Likewise, local harmony (the current chord) can be used to inform note transcription. The converse problem, determining the chord given a set of detected notes, is also a transcription task, which implicitly involves selecting the subset of notes deemed to have a harmonic function. A chord transcription system which uses a probabilistic framework to jointly model the key, metre, chord and bass notes (Mauch and Dixon 2010) gave the best results for the MIREX 2009 and 2010 chord detection tasks. Another example is Raczynski et al. (2010), where the authors introduce a joint probabilistic model for AMT that models the temporal dependencies between musical notes and the underlying chords, as well as the instantaneous dependencies between chords, notes and the observed note saliences based on musicological models encoded in a dynamic Bayesian network. Further examples of joint estimation include the use of graphical model for singing transcription, jointly

estimating pitch, rhythm, segmentation and tempo (Raphael 2005), and an HMM for simultaneous estimation of chords and downbeats (Papadopoulos and Peeters 2011).

Finally, information can also be integrated over time. Most AMT systems model only short-term dependencies, often using Markov models to describe expected melodic, harmonic and rhythmic sequences. Musicological models could be employed to describe these local sequential dependencies (Ryynänen and Klapuri 2005) as well as longer-term relationships such as structural repetition and key. Structural relationships have been utilised, for example, to improve beat tracking (Dannenberg 2005) and chord transcription (Mauch et al. 2009).

6.2 Combining methods targeting the same feature

Information could also be integrated by combining multiple estimators or detectors for a single feature, for instance combining two multi-pitch estimators, especially if these are based on different acoustic cues or different processing principles. This could help overcome weak points in the performance of the individual estimators, offer insight on the weaknesses of each and raise the overall system accuracy provided that the methods do not fail on the same instances. In a different context, several pitched instrument onset detectors, which individually have high precision and low recall, have been successfully combined in order to obtain an improved detection accuracy (Holzapfel et al. 2010). For classification, adaptive boosting (AdaBoost) provides a powerful framework for fusing different classifiers in order to improve the performance (Freund et al. 1999).

6.3 Joint transcription and source separation

Source separation can benefit transcription-related tasks such as multipitch detection and instrument identification. In particular, instrument identification and separation are highly interdependent, and accomplishing one would significantly ease the other, for example by allowing instrument identification to be performed on the separated source signals (Bosch et al. 2012). Another approach is to perform instrument separation and identification jointly, using for example signal model-based probabilistic inference in the score-informed case (Itoyama et al. 2011). Furthermore, ideas and algorithms from the field of source separation can be utilised for AMT, especially regarding the exploitation of spatial information where available (Durrieu and Thiran 2012; Ozerov et al. 2012). On the other hand, music transcription can be used to improve source separation. For example in Fourer and Marchand (2012), the authors propose an informed source separation system where a reference transcription for each instrument signal is computed from separate source signals before mixing and that information is encoded in a watermark embedded in the mixture signal. The information is later used by the decoder in order to separate the source signals from the mixture.

For most music signals, there is only one (mono) or two (stereo) mixture signals available, whereas the number of sources can be much larger. In this case, the separation task is underdetermined, and can only be solved by making some assumptions about the sources. These may include sparsity, non-negativity and independence, or may take the form of structured spectral models like NMF models (Grindlay and Ellis 2011), PLCA models (Bay and Beauchamp 2012), spectral

Gaussian scaled mixture models (Spectral-GSMMs) (Arberet et al. 2012) or the source-filter model for sound production (Heittola et al. 2009). Further constraints such as temporal continuity or harmonicity can be employed together with spectral models. Techniques that employ spectral models for the sources or an NMF-based framework that explicitly models the mixing process have been shown to perform well.

The robustness of source separation can be further improved in the user-assisted scenario. For example in Durrieu and Thiran (2012), the user indicates the desired audio source by selecting the estimated F0 track of that source and subsequently the system refines the selected F0 track, and estimates and separates the relevant source.

7 Creating training data

A large subset of AMT approaches perform experiments only on piano data, e.g. Dessein et al. (2010), Poliner and Ellis (2007), Emiya et al. (2010). One reason is because it is relatively easy to create recordings with aligned ground-truth using e.g. a Disklavier, while there is no straightforward way of creating ground-truth for multiple-instrument music, which typically requires hours of manual transcription and is a challenging (and tedious) task if recordings of separated instrument sources are not available. However, this emphasis on piano music may sometimes lead to models that are suitable for pitched percussive instruments only. Also, most of the current AMT methods involve a training stage, where the parameters of the method are optimised using manually annotated data.

The availability of recorded music with the exact underlying score would open up new and promising opportunities for training complex models, where ground-truth for multiple-instrument recordings is crucial for the further development of sophisticated transcription systems. If musical scores became widely available in high-level digital forms (for example via optical music recognition or crowd-sourced transcriptions), they would provide valuable side-information for signal analysis, and in the limit reduce the transcription task to the alignment of an existing score to the input audio, although it should be noted that different renditions of a piece of music often vary considerably in their instrumentation and arrangement. This would require robust alignment methods which are able to detect transpositions and structural changes in a piece. One example of aligned multiple-instrument transcriptions is the set of syncRWC annotations[1], which were created using the framework described in Ewert et al. (2009). Also, an example of a multi-track dataset for transcription is the Bach10 dataset (Duan et al. 2010), for which ground-truth was creating by performing monophonic pitch tracking on the individual instrument tracks.

Finally, it should be added that large databases alone are not sufficient for creating a robust transcription system. In addition, one needs complex models and good optimization algorithms for training the models. One possible use is to cluster the data (for example, according to automatically detected genres) and then train cluster-specific transcription parameters. When the system is used, the same clustering is applied and the cluster-specific parameters are taken into use.

[1] http://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/SyncRWC/

8 Towards a complete transcription

Most of the aforementioned transcription approaches tackle the problems of multiple-F0 estimation and note onset/offset detection, with a minority of approaches also tacking the estimation of instrument identities, key, or tempo. However, in order to produce output in the form of sheet music, additional issues need to be addressed, such as typesetting, estimation of dynamics, fingering, expressive notation and articulation. Although there are approaches that address many of these individual problems, there exists no 'complete' AMT system to date. A further open problem is how to customise or personalise the system output, for example according to instrument and genre. The need for personalised systems in the greater MIR context is addressed in Goto (2012). As an example for the AMT case, pop music is typically transcribed as a melody and a bass line instead of a complete score (e.g. Goto 2004). Thus, automatic instrument and genre identification could be integrated into an AMT system in order to adapt system parameters (as described in Section 5) and generate style-specific transcriptions.

Regarding typesetting, current tools produce readable scores from MIDI data alone (e.g. Lilypond[2]), however, cues from the music signal could also assist in incorporating additional information into the final score (e.g. expressive features for note phrasing). Another problem which is crucial for typesetting is estimating the meter of music signals, for which a general-purpose system has not yet been developed (Klapuri et al. 2006).

As far as dynamics are concerned, in Ewert and Müller (2011) a method was proposed for estimating note intensities in a score-informed scenario. However, even though the vast majority of AMT techniques can estimate note dynamics using pitch salience or activations, evaluating the performance of current AMT systems for the estimation of note dynamics has not yet been addressed, partly due to a lack of ground truth data. Note also that some additional interpretation of the extracted dynamics over time would need to be made, in order to derive the dynamic markings used in score notation (e.g. crescendo, diminuendo). Most existing ground-truth data does not include note intensities, which are difficult to annotate manually, except for datasets created using reproducing pianos (e.g. Emiya et al. 2010, Poliner and Ellis 2007), which automatically contain relative intensity information such as MIDI note velocities.

For extracting fingering, Kasimi et al. (2007) and Radicioni and Lombardo (2005) proposed dynamic programming-based methods for assigning fingers to notes in polyphonic piano and guitar scores respectively. More recent work (Barbancho et al. 2012) addresses the problem of automatically extracting the fingering configurations for guitar recordings in an AMT framework. Also, the problem of automatic fingering transcription for violin is addressed in Maezawa et al. (2012), by analyzing the input signal to determine the most likely sequence of fingerboard locations along the different strings and string locations. For computing fingering, information from the transcribed signal as well as instrument-specific knowledge is needed. Thus, a robust instrument identification system would need to be incorporated for computing

[2] http://lilypond.org/

fingerings in multi-instrument recordings. There is also still room for refinement in the aforementioned instrument-specific approaches, and current work is focused only on a limited set of instruments, while there is a lack of a systematic evaluation methodology in the fingering extraction problem.

For extracting high-level semantic features relating to expression such as on timing (e.g. rubato) and articulation (e.g. vibrato, legato) some work has been done, mostly in the score-informed case (Scheirer 1997). In Gang et al. (2011) a framework for extracting expressive features both from a score-informed and an uninformed perspective is proposed. For the latter, an AMT system is used prior to the extraction of expressive features which include timing, timbre, pitch deviation, and loudness descriptors. It should be mentioned though that not all of the extracted features directly correspond to expressive notation. Thus, additional work needs to be done in order to provide a mapping between mid-level features and the expressive markings which are used in a music score.

9 Conclusions

Automatic music transcription is a rapidly developing research area where several different approaches are still being actively investigated. There are however several open issues in the AMT problem and the performance of current systems is still not sufficient for certain applications which would require a great degree of accuracy. In this paper we have reviewed the current state of AMT research, identified major challenges, and proposed several promising directions to advance the state of the art. For further insights into current research challenges in MIR, see Serra et al. (2013).

One viable way of advancing the current state of AMT systems is to make use of more information, which could take the form of high-level models reflecting the musical conventions or instrument acoustics relevant to the piece in question, or take advantage of explicit user input in order to select algorithms, set parameters, or resolve ambiguities. For example, we discussed how a genre- or instrument-specific transcription system can utilise high-level models that are more precise and powerful than their more general counterparts. Another promising direction for further research is the combination of multiple processing principles, such as different algorithms with complementary properties which estimate a particular feature, or algorithms which extract various types of musical information, such as the key, metrical structure, and instrument identities, and feed that into a model that provides context for the note detection process. Note detection accuracy is not the only determining factor that enables meaningful end-user applications. Often it is possible to circumvent the limitations of the underlying technology in creative ways. For example in semi-automatic transcription, the problem is redefined as achieving the required transcription accuracy with minimal user effort.

To enable progress in these directions, expertise from a range of disciplines will be needed, such as musicology, acoustics, audio engineering, cognitive science and computing. The infrastructure required to support such multidisciplinary collaboration includes joint conferences, meetings and projects, as well as evaluation frameworks and data sets for comparing competing approaches. It is also important to have end-user applications that drive the development of AMT technology and provide it with

relevant feedback, and in this respect the involvement of industry partners will be crucial. In such a scenario, the sharing of code and data between researchers becomes increasingly important. We discussed how the principles of reproducible research and sustainable software improve the efficiency of research by minimising duplication of effort, allowing researchers to focus more of their time on novel work.

We listed a number of open challenges facing AMT researchers for which no immediate solution is, currently, in sight or they have not been fully addressed by the scientific community yet. However, we believe that AMT research has reached the point where certain practical end-user applications can be built, especially where transcribed notes are used as a basis for extracting higher-level information, and we expect to see many more of these appearing in the near future as the state of AMT research advances.

Acknowledgements E. Benetos is funded by a City University London Research Fellowship. D. Giannoulis and H. Kirchhoff are funded by a Queen Mary University of London CDTA Studentship. We acknowledge the support of the MIReS project, supported by the European Commission, FP7, ICT-2011.1.5 Networked Media and Search Systems, grant agreement No 287711.

References

Abdallah, S.A., & Plumbley, M.D. (2004). Polyphonic transcription by non-negative sparse coding of power spectra. In *5th int. conf. on music information retrieval* (pp. 318–325).

Arberet, S., Ozerov, A., Bimbot, F., Gribonval, R. (2012). A tractable framework for estimating and combining spectral source models for audio source separation. *Signal Processing, 92*(8), 1886–1901.

Barbancho, A., Klapuri, A., Tardon, L., Barbancho, I. (2012). Automatic transcription of guitar chords and fingering from audio. *IEEE Trans. Audio, Speech, and Language Processing, 20*(3), 915–921.

Barbancho, I., de la Bandera, C., Barbancho, A., Tardon, L. (2009). Transcription and expressiveness detection system for violin music. In *Int. conf. audio, speech, and signal processing* (pp. 189–192).

Barbedo, J., & Tzanetakis, G. (2011). Musical instrument classification using individual partials. *IEEE Trans. Audio, Speech, and Language Processing, 19*(1), 111–122.

Bay, M., & Beauchamp, J.W. (2012). Multiple-timbre fundamental frequency tracking using an instrument spectrum library. *The Journal of the Acoustical Society of America, 132*(3), 1886.

Bay, M., Ehmann, A.F., Downie, J.S. (2009). *Evaluation of multiple-F0 estimation and tracking systems*. In *10th int. society for music information retrieval conf.* (pp. 315–320).

Bello, J., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M. (2005). A tutorial on onset detection in musical signals. *IEEE Transactions on Speech and Audio Processing, 13*(5), 1035–1047.

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

Bello, J.P., Daudet, L., Sandler, M.B. (2006). Automatic piano transcription using frequency and time-domain information. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(6), 2242–2251.

Benetos, E., & Dixon, S. (2011). Polyphonic music transcription using note onset and offset detection. In *IEEE international conference on acoustics, speech, and signal processing* (pp. 37–40). Prague, Czech Republic.

Benetos, E., & Dixon, S. (2012). A shift-invariant latent variable model for automatic music transcription. *Computer Music Journal, 36*(4), 81–94.

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., Klapuri, A. (2012). Automatic music transcription: Breaking the glass ceiling. In *13th int. society for music information retrieval conf.* (pp. 379–384).

Benetos, E., Klapuri, A., Dixon, S. (2012). Score-informed transcription for automatic piano tutoring. In *20th European signal processing conf.* (pp. 2153–2157).

Bertin, N., Badeau, R., Richard, G. (2007). Blind signal decompositions for automatic transcription of polyphonic music: NMF and K-SVD on the benchmark. In *IEEE international conference on acoustics, speech, and signal processing* (pp. 65–68).

Bertin, N., Badeau, R., Vincent, E. (2010). Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Trans. Audio, Speech, and Language Processing, 18*(3), 538–549.

Böck, S., Arzt, A., Krebs, F., Schedl, M. (2012). Online realtime onset detection with recurrent neural networks. In *Proceedings of the 15th international conference on digital audio effects*.

Bosch, J., Janer, J., Fuhrmann, F., Herrera, P. (2012). A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *13th int. society for music information retrieval conf.* (pp. 559–564).

Brown, J. (1991). Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America, 89*(1), 425–434.

Buckheit, J.B., & Donoho, D.L. (1995). *WaveLab and reproducible research*. Tech. Rep. 474, Dept of Statistics, Stanford Univ.

Burred, J., Robel, A., Sikora, T. (2009). Polyphonic musical instrument recognition based on a dynamic model of the spectral envelope. In *Int. conf. audio, speech, and signal processing* (pp. 173–176).

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M. (2008). Content-based music information retrieval: current directions and future challenges. *Proceedings of the IEEE, 96*(4), 668–696.

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

Cemgil, A.T. (2004). *Bayesian music transcription*. Ph.D. thesis, Radboud University Nijmegen, Netherlands.

Cemgil, A.T., Kappen, H.J., Barber, D. (2006). A generative model for music transcription. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(2), 679–694.

Collins, N. (2005). A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *118th convention of the audio engineering society*. Barcelona, Spain.

Cont, A. (2006). Realtime multiple pitch observation using sparse non-negative constraints. In *7th international conference on music information retrieval*.

Dannenberg, R. (2005). Toward automated holistic beat tracking, music analysis, and understanding. In *6th int. conf. on music information retrieval* (pp. 366–373).

Davies, M., & Plumbley, M. (2007). Context-dependent beat tracking of musical audio. *IEEE Transactions on Audio, Speech and Language Processing, 15*(3), 1009–1020.

Davy, M., Godsill, S., Idier, J. (2006). Bayesian analysis of western tonal music. *Journal of the Acoustical Society of America, 119*(4), 2498–2517.

Degara, N., Davies, M., Pena, A., Plumbley, M. (2011). Onset event decoding exploiting the rhythmic structure of polyphonic music. *IEEE Journal of Selected Topics in Signal Processing, 5*(6), 1228–1239.

Degara, N., Rua, E.A., Pena, A., Torres-Guijarro, S., Davies, M., Plumbley, M. (2012). Reliability-informed beat tracking of musical signals. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(1), 290–301.

Desain, P., & Honing, H. (1999). Computational models of beat induction: the rule-based approach. *Journal of New Music Research, 28*(1), 29–42.

Dessein, A., Cont, A., Lemaitre, G. (2010). Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *11th int. society for music information retrieval conf.* (pp. 489–494).

Dittmar, C., & Abeßer, J. (2008). Automatic music transcription with user interaction. In *34. Deutsche jahrestagung für akustik (DAGA)* (pp. 567–568).

Dittmar, C., Cano, E., Abeßer, J., Grollmisch, S. (2012). Music information retrieval meets music education. In M. Müller, M. Goto, M. Schedl (Eds.), *Multimodal music processing. Dagstuhl follow-ups* (Vol. 3, pp. 95–120). Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research, 30*(1), 39–58.

Dixon, S., Goebl, W., Cambouropoulos, E. (2006). Perceptual smoothness of tempo in expressively performed music. *Music Perception, 23*(3), 195–214.

Dressler, K. (2012). Multiple fundamental frequency extraction for MIREX 2012. In *Music information retrieval evaluation eXchange*. http:www.music-ir.org/mirex/abstracts/2012/KD1.pdf.

Duan, Z., Pardo, B., Zhang, C. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(8), 2121–2133.

Durrieu, J., & Thiran, J. (2012). Musical audio source separation based on user-selected F0 track. In *10th int. conf. latent variable analysis and source separation* (pp. 438–445).

Eggink, J., & Brown, G. (2003). A missing feature approach to instrument identification in polyphonic music. In *Int. conf. audio, speech, and signal processing* (Vol. 5, pp. 553–556).

Emiya, V., Badeau, R., David, B. (2010). Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(6), 1643–1654.

Ewert, S., & Müller, M. (2011). Estimating note intensities in music recordings. In *Int. conf. audio, speech, and signal processing* (pp. 385–388).

Ewert, S., & Müller, M. (2012). Using score-informed constraints for NMF-based source separation. In *Int. conf. audio, speech, and signal processing* (pp. 129–132).

Ewert, S., Muller, M., Grosche, P. (2009). High resolution audio synchronization using chroma onset features. In *IEEE international conference on audio, speech and signal processing* (pp. 1869–1872).

Eyben, F., Böck, S., Schuller, B., Graves, A. (2012). Universal onset detection with bidirectional long short-term memory neural networks. In *11th international society for music information retrieval conference*.

Fourer, D., & Marchand, S. (2012). Informed multiple-F0 estimation applied to monaural audio source separation. In *20th European signal processing conf.* (pp. 2158–2162).

Freund, Y., Schapire, R., Abe, N. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence, 14*(5), 771–780.

Fuentes, B., Badeau, R., Richard, G. (2011). Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA. In *Int. conf. audio, speech, and signal processing* (pp. 401–404).

Fuentes, B., Badeau, R., Richard, G. (2012). Blind harmonic adaptive decomposition applied to supervised source separation. In *20th European signal processing conf.* (pp. 2654–2658).

Gang, R., Bocko, G., Lundberg, J., Roessner, S., Headlam, D., Bocko, M. (2011). A real-time signal processing framework of musical expressive feature extraction using MATLAB. In *12th int. society for music information retrieval conf.* (pp. 115–120).

Giannoulis, D., & Klapuri, A. (2013). Musical instrument recognition in polyphonic audio using missing feature approach. In *IEEE transactions on audio, speech, and language processing* (Vol. 21, no. 9, pp. 1805–1817). doi:10.1109/TASL.2013.2248720.

Gillet, O., & Richard, G. (2003). Automatic labelling of tabla signals. In *4th int. conf. on music information retrieval*.

Goto, M. (2004). A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication, 43*, 311–329.

Goto, M. (2012). Grand challenges in music information research. In M. Müller, M. Goto, M. Schedl (Eds.), *Multimodal music processing. Dagstuhl follow-ups* (Vol. 3, pp. 217–225). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Goto, M., Hashiguchi, H., Nishimura, T., Oka, R. (2002). RWC music database: Popular, classical, and jazz music databases. In *Proc. ISMIR* (Vol. 2, pp. 287–288).

Gouyon, F., & Dixon, S. (2005). A review of automatic rhythm description systems. *Computer Music Journal, 29*(1), 34–54.

Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C. (2006). An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech and Language Processing, 14*(5), 1832–1844.

Grindlay, G., & Ellis, D. (2011). Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments. *IEEE Journal of Selected Topics in Signal Processing, 5*(6), 1159–1169.

Grosche, P., Schuller, B., Müller, M., Rigoll, G. (2012). Automatic transcription of recorded music. *Acta Acustica United with Acustica, 98*(2), 199–215.

Heittola, T., Klapuri, A., Virtanen, T. (2009). Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *10th int. society for music information retrieval conf.* (pp. 327–332).

Herrera-Boyer, P., Klapuri, A., Davy, M. (2006). Automatic classification of pitched musical instrument sounds. In *Signal processing methods for music transcription* (pp. 163–200).

Holzapfel, A., Stylianou, Y., Gedik, A., Bozkurt, B. (2010). Three dimensions of pitched instrument onset detection. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(6), 1517–1527.

Huang, X., Acero, A., Hon, H.W. (Eds.). (2001). *Spoken language processing: A guide to theory, algorithm and system development*. Prentice Hall.

Humphrey, E.J., Bello, J.P., LeCun, Y. (2013). Feature learning and deep architectures: new directions for music informatics. *Journal of Intelligent Information Systems*. doi:10.1007/s10844-013-0248-5.

Itoyama, K., Goto, M., Komatani, K., Ogata, T., Okuno, H. (2011). Simultaneous processing of sound source separation and musical instrument identification using Bayesian spectral modeling. In *Int. conf. audio, speech, and signal processing* (pp. 3816–3819).

Izmirli, O. (2005). An algorithm for audio key finding. In *Music information retrieval evaluation exchange*. http://www.music-ir.org/mirex/abstracts/2005/izmirli.pdf.

Kameoka, H., Nishimoto, T., Sagayama, S. (2007). A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Transactions on Audio, Speech, and Language Processing, 15*(3), 982–994.

Kameoka, H., Ochiai, K., Nakano, M., Tsuchiya, M., Sagayama, S. (2012). Context-free 2D tree structure model of musical notes for Bayesian modeling of polyphonic spectrograms. In *13th int. society for music information retrieval conf.* (pp. 307–312).

Kasimi, A.A., Nichols, E., Raphael, C. (2007). A simple algorithm for automatic generation of polyphonic piano fingerings. In *8th international conference on music information retrieval* (pp. 355–356). Vienna, Austria.

Kirchhoff, H., Dixon, S., Klapuri, A. (2012). Shift-variant non-negative matrix deconvolution for music transcription. In *Int. conf. audio, speech, and signal processing* (pp. 125–128).

Kitahara, T., Goto, M., Komatani, K., Ogata, T., Okuno, H.G. (2007). Instrogram: probabilistic representation of instrument existence for polyphonic music. *Information and Media Technologies, 2*(1), 279–291.

Klapuri, A. (2003). Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Transactions on Audio, Speech, and Language Processing, 11*(6), 804–816.

Klapuri, A., Davy, M. (Eds.). (2006). *Signal processing methods for music transcription*. Springer.

Klapuri, A., Eronen, A., Astola, J. (2006). Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech, and Language Processing, 14*(1), 342–355.

Klapuri, A., Eronen, A., Seppänen, J., Virtanen, T. (2001). Automatic transcription of music. In *Symposium on stochastic modeling of music*. Ghent, Belgium.

Koretz, A., & Tabrikian, J. (2011). Maximum a posteriori probability multiple pitch tracking using the harmonic model. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(7), 2210–2221.

Lacoste, A., & Eck, D. (2007). A supervised classification algorithm for note onset detection. *EURASIP Journal on Applied Signal Processing, 2007*(1), 1–13. ID 43745.

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

Lee, C.T., Yang, Y.H., Chen, H. (2012). Multipitch estimation of piano music by exemplar-based sparse representation. *IEEE Trans. Multimedia, 14*(3), 608–618.

Lee, K., & Slaney, M. (2008). Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio. *IEEE Transactions on Audio, Speech and Language Processing, 16*(2), 291–301.

Leveau, P., Vincent, E., Richard, G., Daudet, L. (2008). Instrument-specific harmonic atoms for mid-level music representation. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(1), 116–128.

Little, D., & Pardo, B. (2008). Learning musical instruments from mixtures of audio with weak labels. In *9th int. conf. on music information retrieval* (p. 127).

Loscos, A., Wang, Y., Boo, W. (2006). Low level descriptors for automatic violin transcription. In *7th int. conf. on music information retrieval* (pp. 164–167).

Maezawa, A., Itoyama, K., Komatani, K., Ogata, T., Okuno, H.G. (2012). Automated violin fingering transcription through analysis of an audio recording. *Computer Music Journal, 36*(3), 57–72.

Marolt, M. (2012). Automatic transcription of bell chiming recordings. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(3), 844–853.

Mauch, M., & Dixon, S (2010). Simultaneous estimation of chords and musical context from audio. *IEEE Transactions on Audio, Speech, and Language Processing, 18*(6), 1280–1289.

Mauch, M., Noland, K., Dixon, S. (2009). Using musical structure to enhance automatic chord transcription. In *10th int. society for music information retrieval conf.* (pp. 231–236).
McKinney, M., Moelants, D., Davies, M., Klapuri, A. (2007). Evalutation of audio beat tracking and music tempo extraction algorithms. *Journal of New Music Research, 36*(1), 1–16.
Music Information Retrieval Evaluation eXchange (MIREX) (2011). http://music-ir.org/mirexwiki/. Accessed 8 Jul 2013.
Müller, M., Ellis, D., Klapuri, A., Richard, G. (2011). Signal processing for music analysis. *IEEE J. Selected Topics in Signal Processing, 5*(6), 1088–1110.
Nam, J., Ngiam, J., Lee, H., Slaney, M. (2011). A classification-based polyphonic piano transcription approach using learned feature representations. In *12th int. society for music information retrieval conf.* (pp. 175–180).
Nesbit, A., Hollenberg, L., Senyard, A. (2004). Towards automatic transcription of Australian aboriginal music. In *5th int. conf. on music information retrieval* (pp. 326–330).
Noland, K., & Sandler, M. (2006). Key estimation using a hidden markov model. In *Proceedings of the 7th international conference on music information retrieval (ISMIR)* (pp. 121–126).
Ochiai, K., Kameoka, H., Sagayama, S. (2012). Explicit beat structure modeling for non-negative matrix factorization-based multipitch analysis. In *Int. conf. audio, speech, and signal processing* (pp. 133–136).
O'Hanlon, K., Nagano, H., Plumbley, M. (2012). Structured sparsity for automatic music transcription. In *IEEE international conference on audio, speech and signal processing* (pp. 441–444).
Oram, A., & Wilson, G. (2010). *Making software: What really works, and why we believe it.* O'Reilly Media, Incorporated.
Oudre, L., Grenier, Y., Févotte, C. (2009). Template-based chord recognition: Influence of the chord types. In *10th international society for music information retrieval conference* (pp. 153–158).
Özaslan, T., Serra, X., Arcos, J.L. (2012). Characterization of embellishments in Ney performances of Makam music in Turkey. In *13th int. society for music information retrieval conf.*
Ozerov, A., Vincent, E., Bimbot, F. (2012). A general flexible framework for the handling of prior information in audio source separation. *IEEE Trans. Audio, Speech, and Language Processing, 20*(4), 1118–1133.
Papadopoulos, H., & Peeters, G. (2008). Simultaneous estimation of chord progression and downbeats from an audio file. In *IEEE international conference on acoustics, speech and signal processing* (pp. 121–124).
Papadopoulos, H., & Peeters, G. (2011). Joint estimation of chords and downbeats from an audio signal. *IEEE Transactions on Audio, Speech and Language Processing, 19*(1), 138–152.
Peeling, P., & Godsill, S. (2011). Multiple pitch estimation using non-homogeneous Poisson processes. *IEEE J. Selected Topics in Signal Processing 5*(6), 1133–1143.
Peeters, G. (2006). Musical key estimation of audio signal based on hidden Markov modeling of chroma vectors. In *Proceedings of the 9th international conference on digital audio effects* (pp. 127–131).
Pertusa, A., & Iñesta, J.M. (2008). Multiple fundamental frequency estimation using Gaussian smoothness. In *int. conf. audio, speech, and signal processing* (pp. 105–108).
Poliner, G., & Ellis, D. (2007). A discriminative model for polyphonic piano transcription. *EURASIP J. Advances in Signal Processing 8*, 154–162.
Poliner, G., Ellis, D., Ehmann, A., Gomez, E., Streich, S., Ong, B. (2007). Melody transcription from music audio: Approaches and evaluation. *IEEE Trans. Audio, Speech, and Language Processing 15*(4), 1247–1256.
Raczyński, S.A., Ono, N., Sagayama, S. (2009). Note detection with dynamic bayesian networks as a postanalysis step for NMF-based multiple pitch estimation techniques. In *IEEE workshop on applications of signal processing to audio and acoustics* (pp. 49–52).
Raczynski, S.A., Vincent, E., Bimbot, F., Sagayama, S., et al. (2010). Multiple pitch transcription using DBN-based musicological models. In *2010 int. society for music information retrieval conf.* (ISMIR) (pp. 363–368).
Radicioni, D.P., & Lombardo, V. (2005) Fingering for music performance. In *International computer music conference* (pp. 527–530).
Raphael, C. (2005). A graphical model for recognizing sung melodies. In *6th international conference on music information retrieval* (pp. 658–663).
Reis, G., Fonseca, N., de Vega, F.F., Ferreira, A. (2008). Hybrid genetic algorithm based on gene fragment competition for polyphonic music transcription. In *Conf. applications of evolutionary computing* (pp. 305–314).

Röbel, A. (2005). Onset detection in polyphonic signals by means of transient peak classification. In *Music information retrieval evaluation exchange*. http://www.music-ir.org/evaluation/mirex-results/articles/onset/roebel.pdf.

Ryynänen, M., & Klapuri, A. (2005). Polyphonic music transcription using note event modeling. In *IEEE workshop on applications of signal processing to audio and acoustics* (pp. 319–322).

Ryynänen, M., & Klapuri, A. (2008). Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal 32*(3), 72–86.

Scheirer, E. (1997). Using musical knowledge to extract expressive performance information from audio recordings. In H. Okuno, D. Rosenthal (Eds.), *Readings in computational auditory scene analysis*. Lawrence Erlbaum.

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jorda, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., Widmer, G. (2013). *Roadmap for music information research.* Creative Commons BY-NC-ND 3.0 license. http://mires.eecs.qmul.ac.uk.

Smaragdis, P., & Brown, J.C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *IEEE workshop on applications of signal processing to audio and acoustics* (pp. 177–180).

Smaragdis, P., & Mysore, G.J. (2009). Separation by humming: User-guided sound extraction from monophonic mixtures. In *IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*. USA: New Paltz.

Smaragdis, P., Raj, B., Shashanka, M. (2006). A probabilistic latent variable model for acoustic modeling. In *Neural information processing systems workshop.* Canada: Whistler.

Vandewalle, P., Kovacevic, J., Vetterli, M. (2009). Reproducible research in signal processing. *Signal Processing Magazine, IEEE 26*(3), 37–47.

Vincent, E., Bertin, N., Badeau, R. (2010). Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. Audio, Speech, and Language Processing, 18*(3), 528–537.

Wang, Y., & Zhang, B. (2008). Application-specific music transcription for tutoring. *IEEE MultiMedia 15*(3), 70–74.

Wilson, G., Aruliah, D., Brown, C.T., Hong, N.P.C., Davis, M., Guy, R.T., Haddock, S.H., Huff, K., Mitchell, I.M., Plumbley, M.D., et al. (2012). Best practices for scientific computing. arXiv preprint arXiv:1210.0530.

Wu, J., Vincent, E., Raczynski, S., Nishimoto, T., Ono, N., Sagayama, S. (2011). Multipitch estimation by joint modeling of harmonic and transient sounds. In *Int. conf. audio, speech, and signal processing* (pp. 25–28).

Yeh, C. (2008). *Multiple fundamental frequency estimation of polyphonic recordings.* Ph.D. thesis, Université Paris VI - Pierre et Marie Curie, France.

Yoshii, K., & Goto, M. (2012). A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation. *IEEE Trans. Audio, Speech, and Language Processing 20*(3), 717–730.

Zhou, R., & Reiss, J. (2007). Music onset detection combining energy-based and pitch-based approaches. In *Music information retrieval evaluation exchange.* http://www.music-ir.org/mirex/abstracts/2007/OD_zhou.pdf.