J Braz Comput Soc (2013) 19:589–604
DOI 10.1007/s13173-013-0118-6

SURVEY PAPER

Survey on automatic transcription of music

Historical overview of techniques

Tiago Fernandes Tavares · Jayme Garcia Arnal Barbedo · Romis Attux · Amauri Lopes

Received: 11 October 2012 / Accepted: 11 July 2013 / Published online: 6 August 2013

Abstract An automatic music transcriber is a device that detects, without human interference, the musical gestures required to play a particular piece. Many techniques have been proposed to solve the problem of automatic music transcription. This paper presents an overview on the theme, discussing digital signal processing techniques, pattern classification techniques and heuristic assumptions derived from music knowledge that were used to build some of the main systems found in the literature. The paper is focused on the motivations behind each technique, aiming to serve both as an introduction to the theme and as resource for the development of new solutions for automatic transcription.

Keywords Automatic transcription of music · Digital signal processing · Machine learning · Review

1 Introduction

The task of transcribing a musical piece consists of identifying the musical gestures that are required to reproduce it, based on the corresponding acoustic signal. Through transcription, a vector representation of the musical signal is obtained, which allows the analysis of several semantic features related to the acoustic signal. Since the end of the 1970s, many systems for automatic music transcription (AMT) have been proposed. However, no generic and robust solution has yet been obtained.

Applications for AMT are all systems that receive acoustic signals as input, but are more effective if a symbolic representation is provided. Among them, there are some already implemented and usable by end-users, such as query-by-content databases [62] and educational software [17,117,121], but improvements in the field will allow the construction of applications that require a greater accuracy, such as music analysis and documentation devices.

Research on the theme is necessarily multidisciplinary, involving digital signal processing, machine learning and musical models. Therefore, researchers have split the problem into many different sub-problems. The construction of a complete AMT system involves understanding all those sub-problems and properly integrating the required solutions. Also, progress in the development of new solutions requires understanding how the existing methods are related to the mathematical models for auditory phenomena. This means that an overview of the problem, covering not only the techniques themselves, but also the underlying motivations and the context in which they were developed, is of great importance.

Aiming at providing that understanding, this paper presents an overview on automatic transcription of music. The text includes conceptual remarks, discussions and historical overviews on how specific solutions fit the problem and what are their main advantages and drawbacks. It aims to serve as a resource for the development of new solutions for the problem, but may also be used as an introduction to the theme. For more specific technical details, the reader is encouraged to refer to the bibliography or to textbooks such as [57].

This research was funded by CNPq—Conselho Nacional de Pesquisa e Desenvolvimento Cientifico e Tecnológico.

T. Fernandes Tavares (✉) · R. Attux · A. Lopes
School of Electrical and Computer Engineering, University of Campinas, Av. Albert Einstein, 400, Cidade Universitária Zeferino Vaz , P.O. Box 6101, Campinas, SP 13083-852, Brazil
e-mail: tiagoft@gmail.com

J. Garcia Arnal Barbedo
Embrapa Agricultural Informatics, Av. André Tosello, 209,
Barão Geraldo, P.O. Box 6041, Campinas, SP 13083-886, Brazil

This paper focuses on systems designed to work on polyphonic signals, that is, those in which more than one note can be played at the same time. The transcription of monophonic audio has been largely studied, and many efficient solutions for it have been proposed. Research effort on automatic music transcription, nowadays, is directed towards the more difficult problem of transcribing polyphonic audio.

The organization of this paper is as follows. Section 2 brings some remarks on psycho-acoustics, on how basic models are related to simple auditory sensations and on conventions used for music notation. In Sect. 3, digital signal processing techniques for obtaining proper representations of the signal, allowing further classification, are discussed. Section 4 addresses the problem of finding discrete notes using the previously obtained data. Section 5 shows the different transcription problems tackled in the literature, and how their different characteristics are important in order to build an automatic transcriber. The evaluation of automatic transcription systems is approached in Sect. 6. Further discussions are conducted in Sects. 7 and 8 concludes the text.

2 Signals, notation and sensations

An audio signal is a signal that can trigger auditory sensations. Although this definition may involve diverse signals, some restrictions are normally used, based on the limits of the average human hearing. Thus, in general, an audio signal is assumed to be a variation in the air pressure $x(t)$ with frequency components in the range of 20 Hz to 20 kHz. The characteristics of this sound pressure variation may be controlled by a musician, either using their own voice and body or interacting with musical instruments using proper gestures. When the sound pressure variation is a harmonic signal, that is, $x(t)$ is the sum of sinusoidal components whose frequencies are multiples mF of a fundamental frequency ($F0$) F, as in

$$x(t) = \sum_{m=1}^{M} A_m \cos(2\pi m F t + \phi_m), \qquad (1)$$

it triggers an auditory sensation called pitch [51], which allows classifying sounds in a scale that goes from bass to treble [77].

In Western culture, music is traditionally played using a discrete set of pre-defined pitches [51]. Each one of these elements is called a musical note. Although there are different techniques for tuning notes, it is generally accepted, for automatic transcription purposes, that F0 values are drawn from an equal-tempered scale, defined by steps (or semitones) corresponding to a frequency ratio of $\sqrt[12]{2}$. For historical reasons, notes were called A, A# (or Bb),[1] B, C, C# (or Db), D, D# (or Eb), E, F, F# (or Gb), G and G# (or Ab), in a total of 12 different tones, comprising one octave. In that notation, it is possible to refer to the octave of a specific note using a number, e.g., A4 is note A in the fourth octave. As a consequence of the method for the construction of the scale, the fundamental frequency assigned to the note A3 is half the one related to A4 and so on. Conventionally, the note A4 is tuned so its fundamental frequency is 440 Hz, and this allows defining the $F0$s of all other notes by applying the $\sqrt[12]{2}$ ratio.

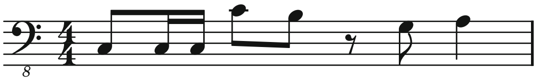

Fig. 1 Example of traditional Western musical score

In order to execute a musical piece, the musician generally follows instructions to play a particular sequence of musical notes. These instructions may be taught by cultural tradition, but written notations were developed across history. Different forms of musical notation have arisen, each one being more adequate to specific forms of playing, understanding and composing music [54].

In the Western culture, a common form of notation is the score, as shown in Fig. 1. In this notation, each note to be played is represented by a symbol (a vertical line and a notehead) in a staff with five lines. Details on the symbol describing each note are used to represent its duration, that is, for how long the note should be played, and the vertical positioning of each notehead describes which note should be played. There are many other characteristics of music that may be written in a musical score, but a complete description is beyond the scope of this paper.

It is important to notice that the score notation presents relative timing, that is, note durations are represented as a ratio of each other, and not as absolute values. This allows a particular piece to be played faster or slower, according to the interpretation of the musician. Less freedom is given to the interpreter by using a notation that is more accurate in time, for example, a piano-roll. In this notation, each note event is represented by its onset and offset (that is, the exact time when the note should start and stop playing), and its pitch. For the purposes of inter-device communication using the MIDI protocol, the pitch of each note is described by an integer (called MIDI number) calculated by:

$$p = 69 + 12 \log_2 \frac{F}{440}. \qquad (2)$$

Because of the extensive use of MIDI devices in the context of electronic music, the piano-roll notation is often referred to as *MIDI notation*. The piano-roll related to a particular piece may be visualized in a Cartesian plane, where the y-axis represents the pitch p and the x-axis represents time. This representation may be seen in Fig. 2.

All kinds of notation have their own advantages and drawbacks. For example, although the piano-roll notation allows

[1] A# is read as "A sharp" and Bb is read as "B flat".

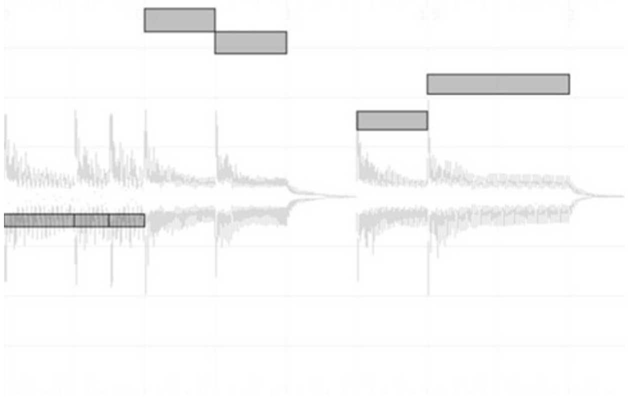

Fig. 2 Example of a piano-roll and the related acoustic signal, which was synthesized from the score in Fig. 1

one to know exactly what is the duration intended for each note, it is not as easily readable or interpretable by a human being as the traditional score.

In order to obtain the transcription of a particular piece, it is necessary to define the symbols that best describe the piece considering the desired notation. When the transcription process is performed by a device without human interference, it is called automatic transcription. Transcription of music is only possible because auditory sensations related to different musical gestures are distinguishable. To build an automatic transcriber, it is necessary to understand the relationship between these sensations and particular signal models, as well as the conditions under which these models work.

2.1 Pitch

The harmonic model in Expression 1 is known to be overly simplistic, as audio synthesis derived directly from that model is easily recognized as artificial. However, it is still used in many audio-related applications. There are several methods to detect the fundamental frequency (and, therefore, the pitch) of a harmonic signal. Hence, if only one note is played at each time instant (in this case, the audio signal is said to be monophonic), the application of the model in Expression 1 is straightforward. When J notes are played concurrently, the resulting signal may be described as

$$y(t) = \sum_{j=1}^{J} \sum_{m=1}^{M_j} A_{m,j} \cos(2\pi m F_j + \phi_{m,j}). \tag{3}$$

The sensation that arises from hearing this signal is that of a sum of sounds with different pitches. This is what happens when more than one note of a piano, for example, is played at the same time.

Both Expressions 1 and 3 can only be used in stationary excerpts, that is, while musical notes are sustained. When transient behavior is found, e.g., during note onsets, different signal characteristics are found and, therefore, different techniques are necessary to detect them, as will be further discussed.

2.2 Onsets

Onsets of new musical notes may be detected by a change on the stationary behavior of the signal. A complete tutorial on the detection of onsets was written by Bello [8]. Note onsets may be detected using typical characteristics of starting notes, which are:

1. Increase in the signal power, for notes with sharp attacks, like a plucked string.
2. Change in the spectral content (that is, the frequency distribution of the signal power), for soft attacks, like a *glissando* in a violin.

These assumptions may be used as an inspiration to build DSP algorithms for finding onsets. These algorithms often rely on the estimation of the progress of the frequency content of the audio signal, which leads to the well-known time-frequency resolution tradeoff [78]. For many years, research focused on obtaining better frequency estimations for short time audio frames, aiming at reaching a better estimation of pitches and onsets. These techniques, called transforms, are discussed in the next section.

3 Transforms

AMT is, generally, performed by digital computers. This means that the acoustic signal is considered a sequence of discrete values. At the same time, it is noticeable that onsets are generally spaced by a fair amount of time—at least some tenths of seconds. Hence, if a random frame of short length is taken from a music signal, it will, with high probability, contain a stationary signal. Therefore, a common approach to AMT is dividing the input signal into frames, operating separately in each frame and then combining the results to obtain the transcription.

There are many different techniques to estimate the frequency components of a specific frame. The inspiration behind those techniques will be discussed below.

The model in Expression 1 can be interpreted as a Fourier series representation. This means that the Fourier transform of the signal $x(t)$ is composed of a series of Dirac delta functions positioned at frequencies mF. The transform, however, is calculated in its discrete form—the DFT—using the sampled signal.[2] $x[n] = x(nt/f_s)$ in a short time frame, as in:

[2] In the sampling expression, f_s is the sampling frequency.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{2\pi jkn}{N}}. \quad (4)$$

The DFT is a linear transform whose magnitude is independent of the phase component of the time-domain signal. Hence, $|X[k]|$ does not depend on any component $\phi_{j,m}$ of the signal in Expression 3. Also, it presents high energy components—lobes—in positions that correspond to the frequencies mF_j. Since the DFT is generally calculated over a short time interval, these lobes have a characteristic bandwidth, which evokes once more the problem of time-frequency resolution.

An early attempt to design an automatic music transcriber using the DFT was made by Piszczalski [87], who noticed that, by analyzing the two highest peaks of the DFT, it is possible to determine the fundamental frequency of a single note in the monophonic case. Later, Privosnik [89] observed that, in Piszczalski's transcription system, there are cases that fail due to the poor frequency resolution. Privosnik [89] speculates that a variable resolution transform, like the Wavelet transform, could be a good solution for this problem.

Keren [55], and later, Hsu and Jang [52], use the multi-resolution Fourier transform (MFT), which consists of calculating magnitude coefficients using different time-domain resolutions. This aims to improve the time-domain resolution for higher frequencies, while the frequency-domain resolution is improved for lower frequencies. Also, the MFT uses prolate-spheroidal functions as bases for the transform, instead of the complex exponentials used in the DFT. Prolate-spheroidal functions have a considerably compact representation in both time and frequency domains [120], which allows a limited time-domain frame to be represented by a low number of coefficients in the frequency domain.

Sterian [106] uses the modal transform (MT) [86], which adaptively modifies the basis function in order to minimize the bandwidth of each lobe, so that $x[n]$ can be represented by fewer coefficients in the frequency domain. Afterwards, Sterian [107] proposed modifications to the MT that improved its potential use in musical applications.

It is important to notice that the DFT may be interpreted as a filter bank. Therefore, a filter bank designed specifically to solve the AMT problem may be potentially functional. The music transcription systems proposed by Moorer [71] relies on filter banks to detect the energy of each harmonic. Miwa [70] uses linear oscillator filters to remove harmonic signals with known fundamental frequencies, so that the filter that eliminates the most energy of the signal indicates the existence of a musical note. Marolt [66,67] uses filter banks with logarithmically-spaced center frequencies, aiming to simulate the behaviour of the human cochlea [81]. This process yields a multi-channel signal that aims to be more correlated to the information that is provided to the human brain. A similar approach is used by Gillet [39].

The idea of using a multi-resolution time-frequency filter bank had been previously proposed by Brown [20], who developed the constant-Q transform (CQT). This transform uses complex exponential bases, similarly to the DFT, but the frame length is different for each coefficient, so that the ratio between the bandwidth of a spectral lobe δf and the frequency f, given by $Q = f/\delta f$, is kept constant. Therefore, to obtain the k-th frequency bin of the transform, it is necessary to use $N[k] = \frac{1}{f_k f_s} Q$ time-domain coefficients. The transform is calculated using a complex exponential basis and each component is normalized:

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} x[n] e^{-\frac{2\pi jQn}{N[k]}}. \quad (5)$$

The CQT was used by Chien [23], and later, by Costantini [24–26], Benetos and Dixon [11] and Kirchhoff et al. [56].

In another attempt to improve the time-frequency resolution, Hainsworth [46] used a technique called time-frequency reassignment (TFR), which was proposed by Kodera [59]. This technique consists of reallocating the energy of a DFT coefficient to a close position in the frequency domain, considering the behavior of three subsequent frames. It allows the resolution of the DFT to be enhanced by improving the precision of the frequency estimation related to each coefficient. However, the TFR technique does not allow the separation of spectral lobes that have not been resolved previously (i.e., merged lobes will not be split), which may worsen its performance in lower frequencies. Later, Barbancho [6] proposed to first estimate the onset and offset of each note and then calculate a lengthier DFT using an adaptive window.

The problem of improving the time-frequency resolution was also addressed by Foo [36], who designed a fast filter bank (FFB) in which the transition between the pass band and the rejected band of each filter is masked by the response of the other filters [63]. The FFB provides a frequency-domain representation for the signal in which spectral lobes are narrower than those yielded by the DFT. Foo [36] observed that by applying the FFB instead of the DFT improved the performance of the AMT system.

In a more recent work, Triki [113] verified that, in acoustic signals, each harmonic partial is modulated in both amplitude and frequency. Therefore, he proposed a frequency domain transform in which the chosen bases may, under some mathematical constraints, change in amplitude and frequency as a function of time. By choosing bases that are more correlated to the input signal, it is possible to obtain spectral representations in which relevant coefficients are more concentrated. Although more accurate, it demands iterative estimation methods, which cause a great loss in speed.

Another possible transform to obtain a frequency domain representation is the chroma spectrum. Used by Oudre

Table 1 Reference table for transforms

References	Technique
[87,89]	Framewise DFT
[39,66,67,70,71,81]	Specialized filter banks for detection—different approaches, all based on the harmonicity principle for pitch generation (Expression 1)
[52,55]	Multi-Resolution Fourier Transform [120]
[106]	Modal Transform [86]
[11,23–26,56]	Constant-Q Transform [20]
[46]	Time-Frequency Reassignment [59]
[36]	Fast Filter Bank [63]
[6]	DFT limited by pre-detected onsets/offsets
[113]	Bases are modulated in amplitude and frequency
[79,80]	Chroma spectrum

[79,80], it consists of a frequency domain spectrum in which there are only 12 coefficients, each corresponding to a specific musical note (octaves are ignored). In this transform, all energy related to the fundamental frequencies corresponding to a note is concentrated in a single coefficient. For example, the coefficient corresponding to A will be the combination of all energy concentrated on 55, 110, 220 Hz, and so on.

The time-frequency resolution tradeoff is an important problem in AMT. Although more accurate models for obtaining frequency-domain representations have been broadly studied, it is noticeable that, lately, the conventional DFT has been preferred over other transforms. That is because the DFT presents some advantages. First, it allows exact reconstruction of the original signal from the transform, which means that all information contained in the time-domain signal is also contained in the frequency-domain representation. Since musical notes are macro events comprising several subsequent DFT analysis windows, it is reasonable to assume that all necessary information to characterize musical notes is present in the DFT magnitude spectrum. Last, the DFT is broadly studied and its behavior is well-known.

Martin [68] states that most errors of his AMT system are due to the existence of harmonically related consonant notes. This is caused by ambiguities related to the sum of harmonic signals, as in Expression 3. In that model, if a partial is found, it is impossible to determine with 100 % accuracy the harmonic series to which it belongs. For that matter, it is necessary to use more complex decision systems, which will use the information from the frequency-domain to find the correct fundamental frequencies of those harmonic series. These will be discussed in the next section.

Table 1 summarizes the information presented in this section, showing what techniques were used to obtain the frequency domain representation of the audio signals analyzed. It is important to note that, although the multi-resolution analysis may prove useful, many recent, state-of-the-art methods rely on a simple framewise DFT analysis.

4 Pattern classification techniques

The final goal of an automatic music transcriber is to obtain a proper description of the musical gestures that have been performed to generate the signal received as input. According to discussions in Sect. 5, that means to infer the pitch, the onset and the offset of each note. The harmonic model in Expression 1 implies that it is necessary to detect what are the frequency partials of a frame of audio in order to infer the pitches of the active notes in that frame. As seen in Sect. 3, there are many different methods to highlight these partials, but, so far, this text has discussed approaches that explicitly decide what are the pitches of the active notes at some time.

The simplest technique to detect partials is based on peak-picking and thresholding in the frequency domain, which corresponds to selecting all local maxima whose absolute value is above a pre-defined threshold [60,68,106,108,113]. A slightly more complex approach involves low-pass filtering the DFT of the analyzed signal so that spurious peaks are eliminated [7]. Another simple operation that may be performed is to apply a parameter estimation algorithm. Hainsworth [45] uses the method proposed by MacLeod [64], which works by obtaining parameters according to a maximum likelihood criterion. In the system proposed by Moorer [71], sub-harmonic spurious components are eliminated by disallowing fundamental tone candidates that do not present even harmonics.

Once the parameters (magnitude and frequency) of the existing partials are found, it is necessary to group them in order to find musical notes. Piszczalski [87] simply assumes that only a single note sounds at each time frame, without any other noise. This is not true in the general case, but may represent some specific, useful cases, like solo flute or voice. Sterian [106] proposes using multiple time frames, grouping partials that do not change in magnitude or frequency by more than a pre-defined threshold. The result of this grouping process is a set of tracks, which are grouped using the harmonicity criterion. Tanaka [108] uses a rule-based system that groups partials considering that they start and end in similar instants, their amplitudes change slowly and they are harmonically related. Lao [60] and Triki [113] add a rule according to which all partials of the same series must have a similar amplitude. Hainsworth [45] separates the tasks of finding notes and classifying them, so that an algorithm based on framewise energy variation finds note onsets and, after that, notes are classified according to the found partials. Similar approaches were also used by Rao and Rao [92] and Uchida and Wada [114]. Finally, Dressler [31] used a chain of pitch formation rules to detect the predominant pitch in an existing mixture.

The process of peak-picking and thresholding is interesting because it filters a great amount of noise and yields few data points. This allows the extraction of information with rule-based systems that rely on psycho-acoustic principles. On the other hand, when a higher number of musical notes are present, this process comprises two phenomena. First, peaks related to partials of different notes will merge, which means that peak-picking becomes more likely to fail. Second, since there may exist a great difference between the loudness of the mixed notes, finding a suitable threshold level will tend to be a harder task.

When the coefficients of a particular frame are interpreted as a vector, the problem of characterizing events in that frame may be considered as a pattern classification problem [33], and many techniques, specifically designed for solving classification problems, may be used.

A pattern classification technique is an algorithm that receives as input a set of characteristics $\boldsymbol{o}$ and yields a label $s(\boldsymbol{o})$ as output. This label characterizes the class to which the object—in this case, the frame—belongs, based on the characteristics $\boldsymbol{o}$. The transcription systems discussed up to this point may be considered rule-based pattern classification systems, which are designed taking into account specialist knowledge. The techniques that will be discussed in the following, however, also gather information from a training dataset.

Artificial neural networks are broadly used classifiers, with the multi-layer perceptron (MLP) being the most emblematic representative [50]. This ANN architecture is based on calculating the function:

$$\boldsymbol{y} = \boldsymbol{A}\begin{bmatrix} f\left(\boldsymbol{B}\begin{bmatrix}\boldsymbol{x}\\ 1\end{bmatrix}\right)\\ 1\end{bmatrix}, \tag{6}$$

where the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, obtained by supervised training, contain weight parameters, $\boldsymbol{x}$ is an input vector, $\boldsymbol{y}$ is an output vector and $f(.)$ is a sigmoidal function [50], like $\tanh(.)$.

The mathematical model in Expression 6 is shown to be capable of universal approximation, that is, it is mathematically capable of modelling any function. Obtaining the parameters that yield the approximation of a particular function depends on a supervised learning algorithm, generally using a gradient descendent approach that iteratively minimizes the approximation error for a given dataset, but often leads to local *minima* [33].

It is important to notice that, in the domain of a transform that presents the characteristics discussed in Sect. 2.1, the magnitude of a particular coefficient may indicate the existence of a frequency component, its non-existence or a state of doubt about its existence. A possible mathematical model for these states is a sigmoidal function $f(x)$. For $x \rightarrow \pm\infty$, the function assumes negative or positive saturation values, which represent either existence or non-existence, and for intermediate values the sigmoid function presents a smooth transition, which represents an increasing value for the hypothesis of existence of the component. MLP networks were used in [9,55,66,67,89], in which one complete network is designed to detect if a particular note is active or inactive; therefore, there are as many networks as possible notes to detect. By using multiple networks, the dimension of the input is considerably reduced, which decreases the susceptibility of the network to converging to local *minima* and having sub-optimal performance.

In order to avoid local *minima*, recent works have preferred to use support vector machines (SVM). A SVM is a machine learning technique that does not minimize an approximation error considering a dataset, but maximizes the distance between the clusters that are to be classified. The training step is unimodal and performed in a single pass (that it, it is not necessary to iterate several times over data, as it is the case for MLP networks). SVM variations were used in [24–26,37–39,88,112,119] to classify musical notes and chords using short time spectral representations, and in [112,118] to classify the timbre of already labeled musical notes.

Recently, deep-belief networks (DBNs) were applied to the problem of transcription by Nam et al. [73]. These networks can be briefly explained as a multi-layer neural network in which each layer is trained independently. DBNs have shown to yield better results than those obtained using SVMs in the databases used by the authors.

Neural networks tend to produce a black box machine, that is, the final transformation obtained may not correspond to the way a specialist would reason over the problem. A more semantically meaningful approach to obtain detection functions is to use probabilistic techniques, which also play an important role in music transcription. Among those techniques, it is important to discuss the Bayesian classifier. This decision technique relates a vector of observations $\boldsymbol{o}$ to the strength of the hypothesis that the observed object belongs to the class s_j, using two factors. The first is the probability that $\boldsymbol{o}$ is generated by using a known model of the class behavior, which gives $P(\boldsymbol{o}|s_j)$. The second is the prior probability of the class s_j. Using the Bayes theorem, the probability of finding the class s_j given the observation vector $\boldsymbol{o}$ is:

$$P(s_j|\boldsymbol{o}) = \frac{P(\boldsymbol{o}|s_j)P(s_j)}{P(\boldsymbol{o})}. \tag{7}$$

Since $P(\boldsymbol{o})$ is equal for all candidate classes s_j, the decision algorithm may discard it and simply choose the class given by $\arg\max_j P(\boldsymbol{o}|s_j)P(s_j)$.

Expression 7 produces a decision system with theoretical minimum probability of error (MPE), which means that Bayesian decision systems may, ideally, reach the best possible classification performance, given the input data.

However, both $P(\boldsymbol{o}|s_j)$ and $P(s_j)$ are unknown and must be estimated, which bounds the performance of the classification systems to the quality of the estimation [33]. Bayesian decision systems were used in [22,41,58,105,111].

When dealing with framewise classification of musical signals, it must be noted that there is a strong correlation between the classes of adjacent frames, because there is a good chance that no new note events happened between these frames. This means that the prior probability of observing a specific class depends on the previous decision.

This premise is used in hidden Markov model (HMM) structures [90]. HMMs are discrete-time systems in which the q-th state depends on the state in the instant $q-1$, that is, $P(s_{j,q}) = P(s_{j,q}|s_{j,q-1})$, which is a property ignored by the Bayesian decision process. Each state may be interpreted as a classification decision, and, with a carefully chosen topology, states will represent the desired musical events. HMMs were used in [37,53,93,96,98,101], with different topologies and data sets.

The same way that probabilistic approaches allow the use of a wide framework of theorems and proven results, another important framework of such kind—the linear algebra—may be used. In this case, a very common model for the detection of musical notes is:

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{A}. \tag{8}$$

In this model, each column $\boldsymbol{x}_q$ must contain a short time frame of the input signal, in a vector representation that does not change significantly if the pitch itself does not change—for example, the absolute value of the DFT of that frame. The representation $\boldsymbol{x}_q$ is factorized as a combination $\boldsymbol{a}_q$ of the basis vectors stored as columns of the $\boldsymbol{B}$ matrix, hence $\boldsymbol{x}_q \approx \boldsymbol{B}\boldsymbol{a}_q$. Therefore, $a_{d,q}$ is the strength of activation of the d-th basis vector during time frame q. Also, the model inherently has a non-negativity constraint, which holds at least for $\boldsymbol{A}$ since it does not make any sense to have a note "subtracted" from the mixture. The representation used in this model is, in general, some kind of spectrogram, hence the values of $\boldsymbol{X}$ and $\boldsymbol{B}$ are also non-negative. Therefore, finding the factors $\boldsymbol{B}$ and $\boldsymbol{A}$ is often referred to as "non-negative matrix factorization" (NMF).

Since the factorization model in Expression 8 is inexact, an approximation must be calculated. This depends on the choice of a suitable error measure. The first one that was used in AMT was the Euclidean distance, that is, the error ϵ is given by:

$$\epsilon = \|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{A}\|. \tag{9}$$

Smaragdis [103] used an iterative gradient-based approach to obtain both $\boldsymbol{B}$ and $\boldsymbol{A}$ from a spectrogram $\boldsymbol{X}$. Although it is a non-supervised learning technique, experiments show that $\boldsymbol{B}$ usually converges to basis vectors corresponding to notes and $\boldsymbol{A}$ converges to their corresponding activation weights. Additionally, Smaragdis [103] observed that note events that are never found separately (notes that are never out of a chord) are not recognized by the system, and, at the same time, some spurious events are recognized. Later, Bertin [13] showed that NMF yields better results than a singular value decomposition (SVD). Sophea [104] and Vincent [115] proposed the use of prior knowledge on the pitches, both restricting the values on the base matrix $\boldsymbol{B}$ so that only values corresponding to the fundamental frequency and the overtones would be allowed to be different than zero. In parallel, Grindlay [43] used instrument data to obtain the base matrix $\boldsymbol{B}$ straight from data, hence it would not have to be updated while searching for the values of $\boldsymbol{A}$.

Bertin [14] converted the usual NMF approach to a probabilistic framework, in which the values of the spectrogram and the factor matrices are interpreted as probabilities. This was useful to connect the factorization results to a hypothesis of temporal smoothness, that is, the expectancy of notes lasting for long time intervals was incorporated to the system. This probabilistic framework is called "Bayesian non-negative matrix factorization" and was later used in experiments for automatic transcription of polyphonic music [15]. A similar technique, namely, probabilistic latent component analysis (PLCA), was used by Han to solve the problem of transcription [48]. Although the calculations performed in the Bayesian approach may be modelled as matrix factorization problems, the use of a probabilistic framework may allow not only a more meaningful interpretation of both the results and the hypothesis over which the system is built on, but also the use of an important set of tools and algorithms, as seen above.

Under the hypothesis that the base matrix $\boldsymbol{B}$ is informed a priori, it is possible to interpret the factorization of each column of $\boldsymbol{X}$, in Expression 9, as an independent problem, that is:

$$\epsilon_q = \|\boldsymbol{x}_q - \boldsymbol{B}\boldsymbol{a}_q\| \tag{10}$$

with the constraint that $a_{q,d} \geq 0, \forall d$, thus obtaining the corresponding weight vector $\boldsymbol{a}_q$.

The hypothesis of time independency allows the construction of causal algorithms for transcription, which is a prerequisite for real-time applications. A gradient-based rule may be used, but an algorithm specifically designed to solve the problem in Expression 10 was proposed by Hanson and Lawson [49]. It was used in the context of AMT by Niedermayer [74], Mauch and Dixon [69] and Tavares et al. [109].

Bertin [16] observed that the Euclidean distance may not be the best error measure for the approximation of Expression 8, using the Itakura-Saito divergence instead, defined as:

$$\epsilon_I(x|y) = \frac{x}{y} - \log\frac{x}{y} - 1. \tag{11}$$

Bertin [16] argues that this divergence could be more suitable for transcription because it is invariant to linear gain (that is, $d(x|y) = d(\lambda x|\lambda y)$).

However, the Itakura–Saito distance is not convex, as opposed to the Euclidean distance. For this reason, Bertin proposes using the β-divergence, given by:

$$\epsilon_\beta(x|y) = \begin{cases} \frac{x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}}{\beta(\beta-1)}, & \beta \in \Re\backslash\{0,1\} \\ x - \log\frac{x}{y} + (x-y), & \beta = 1 \\ \frac{x}{y} - \log\frac{x}{y} - 1, & \beta = 0. \end{cases} \quad (12)$$

Since the β-divergence is convex for $\beta \in [1, 2]$, Bertin [16] proposes minimizing the cost function for $\beta = 2$ and then gradually reducing it until it converges to zero, in which case the β-divergence is equivalent to the Itakura–Saito divergence. The β-divergence was also used by Dessein et al. [29], simply assuming $\beta = 0.5$.

The factorization problem in Expression 8 is close to a source separation problem, in which a number of sources are mixed into different channels. In this case, the sources are the individual notes, each with its own activation level, and the mixture is the final spectrogram. Obtaining the factorization $\boldsymbol{BA}$, under this assumption, is equivalent to solving a blind source separation problem. This interpretation was used by Abdallah [1] and later, by Boulanger–Lewandowsky et al. [19], to build a transcription system based on the assumptions that the activation of the notes is independent and the activation matrix is sparse, that is, few notes are active at each frame. A similar idea was used by Vincent [116], who also incorporated the idea of time dependency for notes and applied the system to recordings, instead of the synthesized harpsichord samples used by Abdallah [1]. The fact that few notes can be played at the same time was exploited by Tavares et al. [109], who applied computer vision techniques to detect which notes can be played in a vibraphone given the position of the mallets, restricting the search space and reducing false positives. Lee et al. [61] incorporated sparseness in the factorization process by minimizing the L1 norm, shown in Expression 13, in the factorization process

$$l_1(\boldsymbol{x}|\boldsymbol{y}) = \sum_{n=1}^{N} \|x_n - y_n\|. \quad (13)$$

Benetos and Dixon [11] observed that, in the constant-Q transform (CQT) domain, templates assigned to different notes are equivalent to shifting other templates in the frequency domain. Later, Kirchoff et al. [56] improved this approach, allowing the use of more than one spectral template per note. The property of shift-invariance was indirectly used by Argenti et al. [4], who used 2D (time-frequency) templates for musical notes, hence considering the spectral changes that happen during the execution of a musical note.

Although factorization methods have shown to be effective, they are computationally expensive and may become prohibitive for large, real-time problems. An inexact solution, however, may be obtained by using the Matching Pursuit algorithm proposed by Mallat and Zhang [65]. This algorithm consists of iteratively selecting, within the dictionary $\boldsymbol{B}$, the vector whose inner product with the input vector is the greatest. The base function is subtracted from the input so that it becomes maximally orthogonal to that base function, and then the algorithm proceeds by analyzing the remainder of that subtraction. The matching pursuit technique, which may be seen as a “greedy” heuristic to minimize Expression 10, may converge to a local minimum, and it will tend to give results that differ from the ones yielded by NNLSQ algorithm when the basis functions are more correlated. It was used for the task of transcription by Derrien [28]. A mixed algorithm developed by O’Hanlon et al. [76] uses a greedy approach to select the notes that are the most efficient to model a mixture, but applies NNLSQ, with a sparsity constraint, to improve the estimation of their ratios.

The decision algorithms described above are those that, historically, have been more frequently used in AMT, but it is also important to mention other relevant techniques. Knowledge-based blackboard systems, like [68], are algorithms that successively incorporate information from multiple, independent agents. Each agent is responsible for dealing with a particular situation, and its output may generate a situation that will be dealt with by other agent. After multiple iterations with all agents, the system yields the final transcription. Reis [94] uses a genetic algorithm, which generates random music, estimates its spectrum and uses the Euclidean distance from the estimated spectrum to the true spectra to evaluate the transcription proposals. Different proposals are combined, changed, replicated and rejected using evolutionary strategies, and after some iterations the desired transcription is obtained. Later, Reis et al. [95] incorporated variance on the spectral envelope and a dynamic noise level analysis, significantly outperforming the previous approach [94].

All these proposals, although using considerably different techniques, are based on similar premises, which were discussed in Sect. 2. These premises are based on physical concepts that, except for a few refinements, have remained the same for a considerably long time [77]. There are, however, particularities of specific transcription problems that must be discussed more thoroughly. This discussion, held in the next section, allows choosing the adequate techniques for each application.

The information presented in this section is condensed in Tables 2 and 3. These tables list, respectively, techniques based on peak-picking and vector classification.

5 An overview of AMT tasks

Human music transcribers are generally specialized in one or a few music genres. This restriction is often brought to AMT research, because it allows the designer to make stronger assumptions on specific signal characteristics, which would be impossible if a generic transcriber were built. In this Section, the most common restrictions for AMT problems will be discussed.

The most common music category that is dealt with in AMT research is the piano solo. The first systems for that task were designed in the 1990s, by Martin [68], Privosnik [89] and Keren [55]. Later, more sophisticated techniques were proposed [2,6,14,16,18,19,24,25,29,44,66,73,76,82, 83,85,88,93,95,103,113,115]. The piano is a polyphonic instrument, that is, more than one note can be played at the same time. Also, there is no direct contact between the musician and the vibrating string that produces the sound (by pressing a key, the musician triggers a mechanical interface), hence a predictable behavior may be expected from the spectral envelope related to each note. Moreover, since it is assumed that there is only one instrument in the acoustic signal, all detected notes should be assigned to it, making timbre analysis unnecessary.

There are also AMT systems that aim to transcribe acoustic signals acquired from an unknown (but unique) instrument. Since timbre is not known a priori, no assumptions can be made regarding the spectral envelope of the detected system. This problem was investigated in 2002, by Chien [23], and after that by [7,60,67,84,96].

The transcription of percussive instruments, also a common problem, involves the detection of unpitched events, that

Table 2 Reference table for peak-picking based classification techniques

References	Technique
Peak detection	
[60,68,106,108,113]	Peak-picking and thresholding
[7]	Peak-picking and thresholding, low-pass filter in frequency domain eliminates spurious notes
[45]	Peak-picking, maximum likelihood criterion [64] is used
[71]	Peak-picking, specialist rule system filters peak hypothesis
Peak grouping	
[87]	Partial grouping with monophonic assumption
[106]	Assumes that partial amplitudes and frequencies do not change significantly across frames
[108]	Partials of same notes begin and end in similar instants
[60,113]	All partials of the same series should have similar amplitudes
[45,92,114]	Notes are classified after onset/offset is detected
[31]	Chain of pitch formation rules highlight predominant pitch

Table 3 Reference table for vector-based pattern classification techniques

References	Technique
[9,55,66,67,89]	Multi-layer perceptron
[24–26,37–39,88,112,119]	Support vector machines
[13–16,26,29,43,75,85,103,104,115]	Non-negative matrix factorization
[73]	Deep-belief networks
[48]	Probabilistic latent variable analysis
[1,19,61,116]	Non-negative matrix factorization with sparsity constrain
[4,11,56]	Shift-invariant factorization
[74]	Non-negative least mean squares
[22,41,58,105,111]	Bayesian decision system
[37,53,93,96,98,101]	Hidden Markov models
[68]	Blackboard system
[28,76]	Matching pursuit
[94,95]	Genetic algorithm

Table 4 Reference table for different problems of transcription

References	Problem
[2,6,14,16,18,19,24,25,29,44,55,66,68,73,76,82,83,85,88,89,93,95,103,113,115]	Piano solo
[7,23,60,67,84,96]	Unknown, unique instrument
[4,7,9,96]	Polyphonic, generic instrument
[43,72]	Polyphonic music with instrument recognition
[3,40,41,45,97–100,109,112]	Specific music genres and/or instruments

is, labelling drum sounds. Such problem, studied by Gillet [37–39], will not be addressed in this paper.

There are also AMT systems that were designed to deal with generic polyphonic audio. Bello [9,7], Ryynanen [96] and Argenti [4] proposed methods capable of dealing with polyphonic audio containing multiple, different instruments, although instruments are not distinguished from each other (i.e., the transcription system assumes that all notes were produced by the same instrument). Muto [72] proposed a system capable of classifying the timbre of notes found in signals containing flutes, pianos, violins and trumpets. The idea of explicitly classifying the timbre of notes is also used by Grindlay [43].

While instrument-related restrictions may give information about how a particular signal behaves, genre-related and instrument-related particularities may provide more robust information regarding how a particular song is composed. AMT systems designed to deal with specific music genres or instruments are found in [3,40,41,45,97–100,109,112].

As stated earlier, there is no AMT system that is, at the same time, generic and robust. The designer of an automatic transcriber must be aware of which musical gestures will be modelled and which will not—for example, if the system should either transcribe the tempo swing of a piano player or not. Such decision must consider the user's needs. In general, most transcription systems are limited to determine which notes are played and their onsets and offsets. This neglects details on specific playing techniques, like pedalling, but, as will be seen, even this simplified version of the generic problem of AMT remains unsolved. The next section conducts a discussion on the evaluation of AMT systems.

Table 4 summarizes the information presented in this section.

6 Evaluation

Evaluating an AMT system means detecting how well it performs, and, furthermore, collect evidence that it may be applied by an end user. The first AMT systems [71,87] were evaluated by visual comparison of ground-truth and automatically obtained scores. This practice remained for some time, but with the growth of scientific efforts towards AMT, it became necessary to use objective performance measures that could not only be automatically calculated over large databases, but also be immediately applied for the performance comparison of different transcribers.

Recall, precision and F-measure are the most frequently used measures. They derive from information retrieval [5] and are used in the music information retrieval exchange (MIREX) [30]. They are defined in Expressions 14, 15 and 16, respectively.

$$\text{Recall} = \frac{\text{\# of correctly transcribed notes}}{\text{\# of notes in ground-truth}}. \tag{14}$$

$$\text{Precision} = \frac{\text{\# of correctly transcribed notes}}{\text{\# of notes in automatic transcription}}. \tag{15}$$

$$\text{F} - \text{measure} = 2\frac{\text{Recall} \times \text{Precision}}{\text{recall} + \text{Precision}}. \tag{16}$$

There is an inherent tradeoff between Recall and Precision that regards the sensitivity of the automatic transcriber. If the sensitivity is too high, many notes will be yielded, the probability of correctly transcribing notes in the ground-truth will be higher, and hence Recall will tend to grow. The increased sensitivity tends to lead the system to yield more false positives, which will cause the Precision to decrease. Conversely, a lower sensitivity parameter tends to lead to higher Precision, with the drawback of lowering the Recall. The F-measure accounts for this tradeoff, representing the harmonic mean between Recall and Precision.

These measures, however, depend on a definition of correctness for a transcribed note. Most modern AMT systems focus on converting audio to a MIDI-like representation in which notes are described by their onsets, offsets and pitches. This allows transcribing tempo fluctuations as played by the musician, but requires a manual annotation of evaluation datasets.

For a yielded note to be considered correct, it is frequently required that its pitch be within half a semitone from the pitch of the corresponding note in the ground-truth. This is justified because deviations less than half a semitone allow obtaining the correct pitch by a simple rounding process. The requirements regarding time, however, are harder to define, as there are many aspects to consider.

The human listening aspect is an important one. An onset difference starts to be perceptually evident above about 10 ms. Onset deviations of more than 20 ms tend to harm the listening experience, but may be used for score following purposes. A deviation over 100 ms is harmful to most applications.

It is also necessary to consider a feasibility aspect. In most AMT systems, the length of the analysis window is between 12 and 43 ms. In live applications, this length is the minimum delay between playing a note and obtaining a response from the system. In offline scenarios, it becomes a timing error, as it impacts in the time-domain analysis resolution.

Last, there is a practical aspect to be considered. It is hard to manually annotate onsets in an audio file below a precision of around some tenths of milliseconds. This aspect was partially solved by using either MIDI synthesizers, MIDI-controlled instruments or by recording musicians that were playing against a MIDI recording.

All of these aspects become even more important when offsets are considered. They are harder to identify (both manually and automatically), and the auditory relevance of their deviations is harder to define numerically. For this reason, many AMT systems ignore offsets and only work with onsets.

Once the timing tolerances are defined, the Recall, Precision and F-measure may be immediately calculated. Although they allow a quick comparison between methods, they show little about the nature of the errors. Furthermore, the impact of the tolerances in the final outcome is not measured, which may generate misleading results.

To account for that, another evaluation method consists in evaluating the transcription in short frames, of around 10 ms. For each frame, the number of active notes in the ground-truth and in the automatic transcription are counted, as well as the number of correctly transcribed notes. These numbers are summed, allowing to calculate Recall, Precision and F-measure. In this case, the measures do not depend on a subjective choice of timing, and long notes are considered proportionally to its duration.

Both the notewise and the framewise evaluation methods, however, fail to yield information regarding the cases in which the transcription algorithm fails. Tavares et al. [110] developed a method that yields histograms for time and pitch deviations. This method highlights information that may be useful, such as detecting the typical time delay, the number of pitch errors for each pitch class and so on. On the other hand, it neglects the need for a unique performance measure.

Daniel and Emiya [27] observed that Recall, Precision and F-measure do not necessarily reflect the perceptual accuracy of a transcription. In reality, different types of errors are perceived differently, for example, timing deviations are less perceptible than pitch deviations. To account for that, Fonseca and Ferreira [35] proposed a perceptually-motivated evaluation method based on applying different evaluation processes to the decaying and sustained parts of a note. This method has shown to be perceptually more accurate than the usual measures.

The standardization of performance measures comprises an important step towards the comparison of AMT systems. However, performance comparison also requires executing different methods over the same dataset. Henceforth, significant effort has been made to provide adequate datasets aiming at future research.

An important step towards test standardization was taken by Goto et al. [42], who developed the Real World Computing database. It comprises audio and MIDI files for 115 popular music songs, 50 classical pieces and 50 jazz pieces. All of them were recorded especially for the database. This database was used by Ryynanen and Klapuri [96–99], Benetos and Dixon [11], Raczynski et al. [91], Simselki and Cemgil [102] and Argenti et al. [4]. However, each one of these works used a different subset of the RWC database.

Aiming at the evaluating transcription of piano solos, Poliner and Elis [88] used a dataset consisting of 92 pieces downloaded from the Classical Piano MIDI Page (http://www.piano-midi.de/) and rendered using standard software (Apple iTunes). A similar dataset was used by Costantini et al. [25], who not only used the same database construction method, but also published an explicit list of all MIDI files used in the process. Nam et al. [73] also used this database, and performed additional tests on it with previous methods [67,96]. The evaluation performed on this database was mostly performed using frame-level note detection. Table 5 shows the average accuracy (F-measure) achieved by each method, as well as a brief description of the techniques employed.

Using synthesized data may lead to scenarios that do not contain aspects such as room reverberation and instrument resonances. The MAPS (standing for MIDI aligned piano sounds) dataset [34] provides high-quality recordings of a Yamaha Disklavier, that is, an automatic piano. It contains 31 GB of recordings with corresponding ground-truths, and may be freely downloaded. It was used by Dessein et al. [29], Nam et al. [73] and O'Hanlon et al. [76]. Table 6 shows the F-measure (both notewise and framewise) reported in each work.

Table 5 Performance comparison for methods using the Poliner–Ellis database

References	Accuracy (%)	Technique
[88]	70	SVM
[25]	85	SVM with memory
[73]	79	DBN
[96]	46	HMM and specialist signal processing
[67]	39	MLP networks

Table 6 Performance comparison for methods using the MAPS database

References	Results		Technique
	Notewise (%)	Framewise (%)	
[29]	71.5	65.5	NMF with *beta*-divergence
[73]	–	74.4	DBN
[67]	–	63.6	MLP networks
[76]	78.2	76.3	Sparse NMF decomposition

Unreported results are identified with a dash

The MIREX quartet [10] is also an important dataset, used as a development set in polyphonic transcription tasks. It contains a recording of the fifth variation of the third movement from Beethoven's String Quartet Op.18 N.5, performed by a woodwind quintet. Each part (flute, clarinet, oboe, bassoon and horn) was recorded separately, and individual annotations are available. This dataset allows a high control on the effects of mixing different instruments together. It was used by Benetos and Dixon [11].

Following a similar idea, Duan et al. [32] developed the MIR-1K dataset, which comprises chorales by J. S. Bach performed by violin, clarinet, saxophone and bassoon. Each part is recorded and annotated separately. This dataset was used by Duan et al. [32] and by Hsu and Jang [52].

Benetos et al. [12] developed and used the score-informed Piano Transcription Dataset, containing seven recordings with intentional execution mistakes. It was designed to be used in tasks in which these mistakes must be detected. However, it can also be used for the evaluation of regular transcription tasks, since the mistakes are carefully annotated.

Further discussion, highlighting some important aspects of the topics cited above, is conducted in the next section.

7 Discussion

Research in AMT has had two major focuses. Most work up to the first half of the decade of 2000 focused on obtaining a transform capable of providing more information about the signal to be analyzed. Afterwards, focus changed to the development of better pattern classification techniques. It is useful to notice that most of the more recent proposals are based on the traditional DFT and, although no clear reason for that was presented, it is reasonable to assume that the facility of interpretation, the broad studies on its properties and the fact that many off-the-shelf computational packages are available are some of the factors that influenced that decision.

The harmonic model that is commonly used to infer pitch (Expression 1), by itself, will lead to ambiguities. Since the fundamental frequencies of simultaneous notes are, frequently, harmonically related, there is a considerable amount of superposition among partials related to each one of these notes [41]. In the presence of noise, either white or due to frequency components that are not part of the model (e.g., the noise of the bow scratching the strings of a violin), it is possible to find multiple hypotheses that may explain the partials found in a specific frame [21,68]. Incorrect estimations resulting from this are common, and no system capable of determining multiple fundamental frequencies without significant flaws has yet been proposed.

It is important to note that there are two main approaches to solve the problem of transcription. The first one is to program sub-systems that apply specialist knowledge aiming to detect psycho-acoustic clues that may lead to the correct transcription, and then combine the results into the desired transcription. If the signal behavior that triggers a particular sensation is known and properly modelled, then its use will imply good results. However, that assumption holds only partially, both because of errors resulting from inexact mathematical modelling and because of errors due to the operation of some algorithms in ill-conditioned situations. Also, as noted by Hainsworth [47], it must be considered that music transcription, when performed by human beings, demands a high level of training and specialization, hence psycho-acoustic models obtained by analysis of non-expert subjects may be ignoring some aspects that are important to build an automatic transcriber. This leads to the second approach to AMT, which is defining a machine learning process that automatically estimates key parameters from a data corpus. These will present all training problems that are typical of machine learning processes, that is, the need for a great amount of data for training, the fact that parameters are not likely to be easily interpretable and the impossibility of theoretically granting that the results are not subject to particularities of a particular database. Machine learning algorithms, however, have been shown to deliver good performance, not rarely outperforming systems built from specialist knowledge.

As it can be seen in Tables 5 and 6, the best transcription results have been obtained when machine learning algorithms are mixed with specialist knowledge. The use of memory for SVMs [25] and sparsity for the NMF [76] lead these systems to outperform previous ones. This indicates that a direction for future AMT research lies on adding specialist knowledge to machine learning processes.

8 Conclusion

This paper has presented an overview on automatic transcription of music. Discussion was conducted so that the two different stages—the digital signal processing, which calculates clues to characterize notes, and the classification process, which ultimately determines note onsets, offsets and pitches—can be properly understood, and what features are desirable in solutions for both stages. It also presents historical remarks on how techniques have evolved in the last 20 years, and the concepts and inspirations behind the most commonly used techniques. The general models that relate physical characteristics of a signal to auditory sensations were also discussed, as well as the specific transcription tasks that are dealt with by most AMT systems.

Although there are many techniques that estimate frequency-domain representations for signals, it was noted that the discrete Fourier transform is the most used. Also, it was noted that techniques that consider long-term characteristics of the signals tend to outperform the others. This suggests that future work in AMT should focus on the development of machine learning techniques that exploit characteristics that are often found in music, such as sparseness in time and frequency and a tendency for continuity in spectral representations.

References

1. Abdallah SA, Plumbley MD (2003) An independent component analysis approach to automatic music transcription. In: Proceedings of 114th AES convention 2003, Amsterdam, The Netherlands
2. Abdallah SA, Plumbley MD (2004) Polyphonic music transcription by non-negative sparse coding of power spectra. In: Proceedings 5th international conference music information retrieval (ISMIR'04), Barcelona, Spain
3. Al-Ghawanmeh F, Jafar IF, A.Al-Taee M, Al-Ghawanmeh MT, Muhsin ZJ (2011) Development of improved automatic music transcription system for the arabian flute (nay). In: Proceedings fo the 8th international multi-conference on systems, signals and devices (SSD), 22–25 Mar 2011
4. Argenti F, Nesi P, Pantaleo G (2011) Automatic transcription of polyphonic music based on the constant-q bispectral analysis. IEEE Trans Audio Speech Lang Process 19(6):1610–1630
5. Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. ACM Press, Addison-Wesley, New York
6. Barbancho I, Barbancho A, Jurado A, Tardon L (2004) Transcription of piano recordings. Appl Acoust 65(12):1261–1287. doi:10.1016/j.apacoust.2004.05.007. http://www.sciencedirect.com/science/article/B6V1S-4D7CDP7-2/
7. Bello J, Daudet L, Sandler M (2006) Automatic piano transcription using frequency and time-domain information. IEEE Trans Audio Speech Lang Process 14(6):2242–2251. doi:10.1109/TASL.2006.872609
8. Bello JP, Daudet L, Abdallah S, Duxbury C, Davies M, Sandler MB (2005) A tutorial on onset detection in music signals. IEEE Trans Audio Speech Lang Process 14(5):1035–1047. doi:10.1109/TASL.2006.872609
9. Bello JP, Monti G, Sandler M, S, M.: Techniques for automatic music transcription. In: Proceedings of the international symposium on music, information retrieval (ISMIR-00), Plymouth, MA, USA, Oct 2000, pp 23–25
10. Benetos E, Dixon S (2011) Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription. IEEE J Sel Topics Signal Process 5(6):1111–1123
11. Benetos E, Dixon S (2011) Multiple-instrument polyphonic music transcription using a convolutive probabilistic model. In: Sound and music computing (SMC 2011)
12. Benetos E, Klapuri A, Dixon S (2012) Score-informed transcription for automatic piano tutoring. In: Proceedings of the 20th European signal processing conference (EUSIPCO 2012)
13. Bertin N, Badeau R, Richard G (2007) Blind signal decompositions for automatic transcription of polyphonic music: Nmf and k-svd on the benchmark. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, ICASSP 2007, vol 1, pp I65–I68. doi:10.1109/ICASSP.2007.366617
14. Bertin N, Badeau R, Vincent E (2009) Fast Bayesian nmf algorithms enforcing harmonicity and temporal continuity in polyphonic music transcription. In: IEEE workshop on applications of signal processing to audio and acoustics, WASPAA '09, pp 29–32. doi:10.1109/ASPAA.2009.5346531
15. 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 Lang Process 18(3):538–549. doi:10.1109/TASL.2010.2041381
16. Bertin N, Fevotte C, Badeau R (2009) A tempering approach for itakura-saito non-negative matrix factorization with application to music transcription. In: IEEE international conference on acoustics, speech and signal processing, ICASSP 2009, pp 1545–1548 (2009). doi:10.1109/ICASSP.2009.4959891
17. Boo W, Wang Y, Loscos A (2006) A violin music transcriber for personalized learning. In: Proceedings of the IEEE international conference on multimedia and expo 2006, pp 2081–2084. doi:10.1109/ICME.2006.262644
18. Boogaart C, Lienhart R (2009) Note onset detection for the transcription of polyphonic piano music. In: Proceedings of the IEEE international conference on multimedia and expo, ICME 2009, pp 446–449. doi:10.1109/ICME.2009.5202530
19. Boulanger-Lewandowski N, Bengio Y, Vincent P (2012) Discriminative non-negative matrix factorization for multiple pitch estimation. Proceedings of the ISMIR 2012, Porto, Portugal
20. Brown JC (1991) Calculation of a constant q spectral transform. J Acoust Soc Am 89(1):425–434
21. Bruno I, Monni S, Nesi P (2003) Automatic music transcription supporting different instruments. In: Proceedings of the 3rd international conference on web delivering of music, 2003 WEDELMUSIC, pp 37–44. doi:10.1109/WDM.2003.1233871
22. Cemgil A, Kappen B, Barber D (2003) Generative model based polyphonic music transcription. In: Proceedings of the 2003 IEEE workshop on applications of signal processing to audio and acoustics, pp 181–184. doi:10.1109/ASPAA.2003.1285861
23. Chien YR, Jeng SK (2002) An automatic transcription system with octave detection. In: Proceedings of the 2002 IEEE international conference on acoustics, speech, and signal (ICASSP), vol 2, pp II–1865.
24. Costantini G, Perfetti R, Todisco M (2009) Event based transcription system for polyphonic piano music. Signal Process 89(9): 1798–1811 (2009). doi:10.1016/j.sigpro.2009.03.024. http://www.sciencedirect.com/science/article/B6V18-4W0R0H7-2/
25. Costantini G, Todisco M, Perfetti R (2009) On the use of memory for detecting musical notes in polyphonic piano music. In: Proceedings of the European conference on circuit theory and design, ECCTD 2009, pp 806–809. doi:10.1109/ECCTD.2009.5275106
26. Costantini G, Todisco M, Perfetti R, Basili R, Casali D (2010) Svm based transcription system with short-term memory oriented to polyphonic piano music. In: Proceedings of the 15th IEEE Mediterranean electrotechnical conference, MELECON 2010–2010, pp 196–201. doi:10.1109/MELCON.2010.5476305
27. Daniel A, Emiya V (2008) Perceptually-based evaluation of the errors usually made when automatically transcribing music. In: Proceedings of the ISMIR 2008, Philadelphia, PA
28. Derrien, O.: Multi-scale frame-based analysis of audio signals for musical transcription using a dictionary of chromatic waveforms. In: Proceedings 2006 IEEE international conference on acoustics,

speech and signal processing, ICASSP 2006, vol 5, p V. doi:10.1109/ICASSP.2006.1661211

29. Dessein A, Cont A, Lemaitre G (2010) Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In: Proceedings of the 11th international Society for Music Information Retrieval conference (ISMIR 2010), Utrecht, Netherlands
30. Downie JS (2006) The music information retrieval evaluation exchange (mirex). D-Lib Magaz 12(12)
31. Dressler K (2011) Pitch estimation by the pair-wise evaluation of spectral peaks. In: Proceedings of the AES 42nd international conference, Ilmenau, Germany
32. Duan Z, Pardo B, Zhang C (2010) Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. IEEE Trans Audio Speech Lang Process 18(8):2121–2133
33. Duda RO, Hart PE, Stork DG (2000) Pattern classification, 2nd edn. Wiley-Interscience, New York
34. Emiya V, Badeau R, David B (2010) Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Trans Audio Speech Lang Process 18(6):1643–1654
35. Fonseca N, Ferreira A (2009) Measuring music transcription results based on a hybrid decay/sustain evaluation. In: Proceedings of the 7th Triennial conference of European Society for the cognitive sciences of music (ESCOM 2009), 12–16 Aug, Jyväskylä, Finland
36. Foo SW, Lee EWT (2002) Transcription of polyphonic signals using fast filter bank. In: Proceedings of the IEEE International Symposium on circuits and systems, ISCAS 2002, vol 3, pp III-241–III-244. dooi:10.1109/ISCAS.2002.1010205
37. Gillet O, Richard G (2004) Automatic transcription of drum loops. In: Proceedings IEEE international conference on acoustics, speech, and signal processing (ICASSP '04), vol 4, pp iv-269–iv-272. doi:10.1109/ICASSP.2004.1326815
38. Gillet O, Richard G (2005) Automatic transcription of drum sequences using audiovisual features. In: Proceedings IEEE international conference on acoustics, speech, and signal processing (ICASSP '05), vol 3, pp iii-205–iii-208. doi:10.1109/ICASSP.2005.1415682
39. Gillet O, Richard G (2008) Transcription and separation of drum signals from polyphonic music. IEEE Trans Audio Speech Lang Process 16(3):529–540. doi:10.1109/TASL.2007.914120
40. Gomez E, Canadas F, Salamon J, Bonada J, Vera P, Cabanas P (2012) Predominant fundamental frequency estimation vs singing voice separation for the automatic transcription of accompanied flamenco singing. In: Proceedings of the 13th international society for music information retrieval conference (ISMIR), 8–12 Oct, Porto, Portugal
41. 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 Commun 43(4):311–329. doi:10.1016/j.specom.2004.07.001. http://www.sciencedirect.com/science/article/B6V1C-4D07TBJ-6/
42. Goto M, Hashiguchi H, Nishimura T, Oka R (2002) Rwc music database: Popular, classical, and jazz music databases. In: Proceedings of the 3rd international conference on music information retrieval (ISMIR 2002), Oct 2002, pp 287–288
43. Grindlay G, Ellis D (2009) Multi-voice polyphonic music transcription using eigeninstruments. In: Proceedinngs of the IEEE workshop on applications of signal processing to audio and acoustics, WASPAA '09, pp 53–56. doi:10.1109/ASPAA.2009.5346514
44. Guibin Z, Sheng L (2007) Automatic transcription method for polyphonic music based on adaptive comb filter and neural network. In: Proceedings of the international conference on mechatronics and automation, ICMA 2007, pp 2592–2597. doi:10.1109/ICMA.2007.4303965
45. Hainsworth S, Macleod MD (2001) Automatic bass line transcription from polyphonic music. In: Proceedings of the international computer music conference, Havana
46. Hainsworth S, Macleod MD, Wolfe PJ (2001) Analysis of reassigned spectrograms for musical transcription. In: Proceedings of the IEEE workshop on applications of signal processing to audio and acoustics, Mohonk Mountain Resort, NY
47. Hainsworth SW, Macleod MD (2007) The automated music transcription problem. Cambridge University Engineering Department, Cambridge
48. Han J, Chen CW (2011) Improving melody extraction using probabilistic latent component analysis. In: Proceedings of the ICASSP 2011, pp 33–36
49. Hanson RJ (1995) Lawson. Solving least squares problems. Philadelphia, CL
50. Haykin S (2000) Neural networks: a comprehensive foundation, 2nd edn. Pearson Education, Prentice Hall
51. Helmholtz H (1885) On the sensation of tone, 4th edn. Dover Publications Inc., New York
52. Hsu CL, Jang JSR (2010) Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion. In: Proceedings of the 11th international society for music information retrieval conference (ISMIR 2010), Utrecht, Netherlands
53. Ning Jiang D, Picheny M, Qin Y (2007) Voice-melody transcription under a speech recognition framework. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, ICASSP 2007, vol 4, pp IV-617–IV-620. doi:10.1109/ICASSP.2007.366988
54. Karkoschka E (1972) Notation in new music: a critical guide to interpretation and realisation. Praeger. http://books.google.ca/books?id=O4MYAQAAIAAJ
55. Keren R, Zeevi YY, Chazan D (1998) Multiresolution time-frequency analysis of polyphonic music. In: Proceedings of the IEEE-SP international symposium on time-frequency and time-scale analysis, pp 565–568, Pittsburgh, PA, USA
56. Kirchhoff H, Dixon S, Klapuri A (2012) Multi-template shift-variant non-negative matrix deconvolution for semi-automatic music transcription. In: Proceedings of the 13th international conference on music information retrieval (ISMIR), Porto, Portugal
57. Klapuri A, Davy M (2006) Signal processing methods for music transcription. Springer, Berlin
58. Kobzantsev A, Chazan D, Zeevi Y (2005) Automatic transcription of piano polyphonic music. In: Proceedings of the 4th international symposium on image and signal processing and analysis, ISPA 2005, pp 414–418. doi:10.1109/ISPA.2005.195447
59. Kodera K, Gendrin R, Villedary C (1978) Analysis of time-varying signals with small bt values. IEEE Trans Acoust Speech Signal Process 26(1):64–76. doi:10.1109/TASSP.1978.1163047
60. Lao W, Tan ET, Kam A (2004) Computationally inexpensive and effective scheme for automatic transcription of polyphonic music. In: Proceedings of the 2004 IEEE international conference on multimedia and expo, ICME '04, vol 3, pp 1775–1778. doi:10.1109/ICME.2004.1394599
61. Lee CT, Yang YH, Chen H (2011) Automatic transcription of piano music by sparse representation of magnitude spectra. In: Proceedings of the 2011 IEEE international conference on multimedia and expo (ICME), pp 1–6. doi:10.1109/ICME.2011.6012000
62. Li J, Han J, Shi Z, Li J (2010) An efficient approach to humming transcription for query-by-humming system. In: Proceedings of the 3rd international congress on image and signal processing (CISP 2010), vol 8, pp 3746–3749. doi:10.1109/CISP.2010.5646801

63. Lim Y (1986) Frequency-response masking approach for the synthesis of sharp linear phase digital filters. IEEE Trans Circ Syst 33(4):357–364. doi:10.1109/TCS.1986.1085930
64. Macleod M (1998) Fast nearly ml estimation of the parameters of real or complex single tones or resolved multiple tones. IEEE Trans Signal Process 46(1):141–148. doi:10.1109/78.651200
65. Mallat S, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. IEEE Trans Signal Process 41(12):3397–3415. doi:10.1109/78.258082
66. Marolt M (2000) Transcription of polyphonic piano music with neural networks. In: Proceedings of the 10th Mediterranean electrotechnical conference, MEleCon 2000, vol 11, pp 512–515
67. Marolt M (2004) A connectionist approach to automatic transcription of polyphonic piano music. IEEE Trans Multimed 6(3): 439–449. doi:10.1109/TMM.2004.827507
68. Martin KD (1996) A blackboard system for automatic transcription of simple polyphonic music. Technical report
69. Mauch M, Dixon S (2010) Approximate note transcription for the improved identification of difficult chords. In: Proceedings of the 11th international society for music information retrieval conference (ISMIR 2010), Utrecht, Netherlands
70. Miwa T, Tadokoro Y, Saito T (1999) Musical pitch estimation and discrimination of musical instruments using comb filters for transcription. In: Proceedings of the 42nd Midwest symposium on circuits and systems, 1999, vol 1, pp 105–108
71. Moorer JA (1977) On the transcription of musical sound by computer. Comput Music J 1(4):32–38
72. Muto Y, Tanaka T (2002) Transcription system for music by two instruments. In: Proceedings of the 6th international conference on signal processing, vol 2, pp 1676–1679. doi:10.1109/ICOSP.2002.1180123
73. Nam J, Ngiam J, Lee H, Slaney M (2011) A classification-based polyphonic piano transcription approach using learned feature representations. In: Proceedings of the 12th international society for music information retrieval conference (ISMIR 2011), 24–28 Oct 2011, Miami, FL, USA
74. Niedermayer B (2008) Non-negative matrix division for the automatic transcription of polyphonic music. In: Proceedings of the ISMIR, pp 544–549
75. O'Grady PD, Rickard ST (2009) Automatic hexaphonic guitar transcription using non-negative constraints. In: Proceedings of signals and systems conference (ISSC 2009), IET Irish, pp 1–6. doi:10.1049/cp.2009.1699
76. O'Hanlon K, Nagano H, Plumbley M (2012) Structured sparsity for automatic music transcription. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 441–444. doi:10.1109/ICASSP.2012.6287911
77. Olson HF (1967) Music, physics and engineering, 2nd edn. Dover Publications Inc., New York
78. Oppenheim AV, Schafer R (1975) Digital signal processing. Prentice-Hall international editions. Prentice-Hall. http://books.google.ca/books?id=vfdSAAAAMAAJ
79. Oudre L, Grenier Y, Fevotte C (2009) Chord recognition using measures of fit, chord templates and filtering methods. In: Proceedings of the IEEE workshop on applications of signal processing to audio and acoustics, WASPAA '09, pp 9–12. doi:10.1109/ASPAA.2009.5346546
80. Oudre L, Grenier Y, Fevotte C (2011) Chord recognition by fitting rescaled chroma vectors to chord templates. In: Processings of the IEEE transactions on audio, speech, and language, vol 17(7):2222–2233. doi:10.1109/TASL.2011.2139205
81. Patterson R, Robinson K, Holdsworth J, McKeown D, Allerhand C (1992) Auditory Physiikigy und perception, chap. complex sounds and auditory images, Exford
82. Peeling P, Cemgil A, Godsill S (2008) Bayesian hierarchical models and inference for musical audio processing. In: Proceedings of the 3rd international symposium on wireless pervasive computing, ISWPC 2008, pp 278–282. doi:10.1109/ISWPC.2008.4556214
83. Peeling P, Cemgil A, Godsill S (2010) Generative spectrogram factorization models for polyphonic piano transcription. IEEE Trans Audio Speech Lang Process 18(3):519–527. doi:10.1109/TASL.2009.2029769
84. Pertusa A, IÃśesta JM (2005) Polyphonic monotimbral music transcription using dynamic networks. Pattern Recogn Lett 26(12):1809–1818. doi:10.1016/j.patrec.2005.03.001. http://www.sciencedirect.com/science/article/B6V15-4FY3NWX-C/
85. Phon-Amnuaisuk S (2010) Transcribing bach chorales using non-negative matrix factorisation. In: Proceedings of the 2010 international conference on audio language and image processing (ICALIP), pp 688–693. doi:10.1109/ICALIP.2010.5685059
86. Pielemeier WJ, Wakefield GH (1996) A high-resolution time-frequency representation for musical instrument signals. J Acoust Soc Am 99(4):2382–2396
87. Piszczalski M, Galler BA (1977) Automatic music transcription. Comput Music J 4(1):24–31
88. Poliner GE, Ellis DP (2007) Improving generalization for classification-based polyphonic piano transcription. In: Proceedings of the 2007 IEEE workshop on applications of signal processing to audio and acoustics, pp 86–89. doi:10.1109/ASPAA.2007.4393050
89. Privosnik M, Marolt M (1998) A system for automatic transcription of music based on multiple agents architecture. In: Proceedings of MELECON'98, pp 169–172 (Tel Aviv 1998)
90. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77(2):257–286
91. Raczynski SA, Vincent E, Bimbot F, Sagayama S (2010) Multiple pitch transcription using dbn-based musicological models. In: Proceedings of the 11th international society for music information retrieval conference (ISMIR 2010), Utrecht, Netherlands
92. Rao V, Rao P (2010) Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. IEEE Trans Audio Speech Lang Process 18(8):2145–2154. doi:10.1109/TASL.2010.2042124
93. Raphael C (2002) Automatic transcription of piano music. In: Proceedings of the 3rd international conference on music information retrieval: ISMIR 2002, pp 15–19, Paris, France
94. Reis G, Fonseca N, Ferndandez F (2007) Genetic algorithm approach to polyphonic music transcription. In: Proceedings of the IEEE international symposium on intelligent signal processing, WISP 2007, pp 1–6. doi:10.1109/WISP.2007.4447608
95. Reis G, Fernandez de Vega F, Ferreira A (2012) Automatic transcription of polyphonic piano music using genetic algorithms, adaptive spectral envelope modeling, and dynamic noise level estimation. IEEE Trans Audio Speech Lang Process 20(8):2313–2328. doi:10.1109/TASL.2012.2201475
96. Ryynanen M, Klapuri A (2005) Polyphonic music transcription using note event modeling. In: Proceedings of the IEEE workshop on applications of signal processing to audio and acoustics, pp 319–322. doi:10.1109/ASPAA.2005.1540233
97. Ryynanen M, Klapuri A (2006) Transcription of the singing melody in polyphonic music. In: Proceedings of the 7th international conference on music information retrieval, Victoria, BC, Canada, pp 222–227
98. Ryynanen M, Klapuri A (2007) Automatic bass line transcription from streaming polyphonic audio. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, ICASSP 2007, vol 4, pp. IV-1437–IV-1440. doi:10.1109/ICASSP.2007.367350
99. Ryynanen M, Klapuri A (2008) Automatic transcription of melody, bass line, and chords in polyphonic music. Comput Music J 32(3):72–86

100. Salamon J, Gulati S, Serra X (2012) A multipitch approach to tonic identification in indian classical music. In: Proceedings of the 13th international society for music information retrieval conference of the (ISMIR), Porto, Portugal
101. Shih HH, Narayanan S, Kuo CC (2002) An hmm-based approach to humming transcription. In: Proceedings of the 2002 IEEE international conference on multimedia and expo, ICME '02, vol 1, pp 337–340. doi:10.1109/ICME.2002.1035787
102. Simsekli U, Cemgil AT (2010) A comparison of probabilistic models for online pitch tracking. In: Proceedings of the 7th conference on sound and music computing (SMC), Barcelona, Spain
103. Smaragdis P, Brown J (2003) Non-negative matrix factorization for polyphonic music transcription. In: Proceedings of the 2003 IEEE workshop on applications of signal processing to audio and acoustics, pp 177–180. doi:10.1109/ASPAA.2003.1285860
104. Sophea S., Phon-Amnuaisuk S (2007) Determining a suitable desired factors for nonnegative matrix factorization in polyphonic music transcription. In: Proceedings of the international symposium on information technology convergence, ISITC 2007, pp 166–170. doi:10.1109/ISITC.2007.50
105. Sterian A, Simoni MH, Wakefield GH (1999) Model-based musical transcription. In: Proceedings of the international computer music conference, Beijing, China
106. Sterian A, Wakefield GH (1996) Robust automated music transcription systems. In: Proceedings of the international computer music conference, Hong Kong
107. Sterian A, Wakefield GH (1997) A frequency-dependent bilinear time-frequency distribution for improved event detection. In: Proceedings of the international computer music conference, Thessaloniki, Greece
108. Tanaka T, Tagami Y (2002) Automatic midi data making from music wave data performed by 2 instruments using blind signal separation. In: Proceedings of the 41st SICE annual conference SICE 2002, vol 1, pp 451–456. doi:10.1109/SICE.2002.1195442
109. Tavares T, Odowichuck G, Zehtabi S, Tzanetakis G (2012) Audio-visual vibraphone transcription in real time. In: Proceedings of the IEEE 14th international workshop on multimedia signal processing (MMSP), pp 215–220. doi:10.1109/MMSP.2012.6343443
110. Tavares TF, Barbedo JGA, Lopes A (2008) Towards the evaluation of automatic transcription of music. In: Proceedings of the VI Brazilian congress of audio, engineering (AES2008), Sao Paulo, Brazil
111. Thornburg H, Leistikow R, Berger J (2007) Melody extraction and musical onset detection via probabilistic models of framewise stft peak data. IEEE Trans Audio Speech Lang Process 15(4):1257–1272. doi:10.1109/TASL.2006.889801
112. Tjahyanto A, Suprapto Y, Purnomo M, Wulandari D (2012) Fft-based features selection for javanese music note and instrument identification using support vector machines. In: Proceedings of the 2012 IEEE international conference on computer science and automation engineering (CSAE), vol 1, pp 439–443. doi:10.1109/CSAE.2012.6272633
113. Triki M, Slock D (2009) Perceptually motivated quasi-periodic signal selection for polyphonic music transcription. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, ICASSP 2009, pp 305–308. doi:10.1109/ICASSP.2009.4959581
114. Uchida Y, Wada S (2011) Melody and bass line estimation method using audio feature database. In: Proceedins of the 2011 IEEE international conference on signal processing, communications and computing (ICSPCC), pp 1–6. doi:10.1109/ICSPCC.2011.6061662
115. Vincent E, Berlin N, Badeau R (2008) Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing, ICASSP 2008, pp 109–112. doi:10.1109/ICASSP.2008.4517558
116. Vincent E, Rodet X (2004) Music transcription with ISA and HMM. In: Proceedings of the 5th international conference on independent component analysis and blind signal separation (ICA), Granada, Espagne, pp 1197–1204. http://hal.inria.fr/inria-00544697
117. Wang Y, Zhang B, Schleusing O (2007) Educational violin transcription by fusing multimedia streams. In: Proceedings of the international workshop on Educational multimedia and multimedia education, Emme '07, ACM, New York, NY, USA, pp 57–66. doi:10.1145/1290144.1290154. http://doi.acm.org/10.1145/1290144.1290154
118. Wang YS, Hu TY, Jeng SK (2010) Automatic transcription for music with two timbres from monaural sound source. In: Proceedings of the 2010 IEEE international symposium on multimedia (ISM), pp 314–317. doi:10.1109/ISM.2010.54
119. Weller A, Ellis D, Jebara T (2009) Structured prediction models for chord transcription of music audio. In: Proceedings of the 2009 international conference on machine learning and applications, ICMLA '09, pp 590–595. doi:10.1109/ICMLA.2009.132
120. Wilson R (1987) Finite prolate spheroidal sequences and their applications i: generation and properties. IEEE Trans Pattern Anal Mach Intell 9(6):787
121. Yin J, Wang Y, Hsu D (2005) Digital violin tutor: an integrated system for beginning violin learners. In: Proceedings of the 13th annual ACM international conference on Multimedia, MULTIMEDIA '05, pp 976–985. ACM, New York. doi:10.1145/1101149.1101353. http://doi.acm.org/10.1145/1101149.1101353