Emmanouil Benetos, Simon Dixon,
Zhiyao Duan, and Sebastian Ewert

Automatic Music Transcription

An overview

©ISTOCKPHOTO.COM/TRAFFIC_ANALYZER

Digital Object Identifier 10.1109/MSP.2018.2869928
Date of publication: 24 December 2018

The capability of transcribing music audio into music notation is a fascinating example of human intelligence. It involves perception (analyzing complex auditory scenes), cognition (recognizing musical objects), knowledge representation (forming musical structures), and inference (testing alternative hypotheses). Automatic music transcription (AMT), i.e., the design of computational algorithms to convert acoustic music signals into some form of music notation, is a challenging task in signal processing and artificial intelligence. It comprises several subtasks, including multipitch estimation (MPE), onset and offset detection, instrument recognition, beat and rhythm tracking, interpretation of expressive timing and dynamics, and score typesetting.

Given the number of subtasks it comprises and its wide application range, it is considered a fundamental problem in the fields of music signal processing and music information retrieval [1], [2]. Because of the very nature of music signals, which often contain several sound sources (e.g., musical instruments and voice) that produce one or more concurrent sound events (e.g., notes and percussive sounds) that are meant to be highly correlated over both time and frequency, AMT is still considered a challenging and open problem in the literature, particularly for music containing multiple instruments and many simultaneous notes (called *polyphonic music* in the music signal processing literature) [2].

The typical data representations used in an AMT system are illustrated in Figure 1. Usually, an AMT system takes an audio waveform as input [Figure 1(a)], computes a time–frequency representation [Figure 1(b)], and outputs a representation of pitches over time [also called a *piano-roll* representation, Figure 1(c)] or a typeset music score [Figure 1(d)].

In this article, we provide a high-level overview of AMT, emphasizing the intellectual merits and broader impacts of this topic and linking AMT to other problems found in the wider field of digital signal processing. We give an overview of approaches to AMT, detailing the methodology used in the two main families of methods, based respectively on deep learning and nonnegative matrix factorization (NMF). Finally, we provide an

 1053-5888/19©2019IEEE

extensive discussion of open challenges for AMT. Regarding the scope of the article, we emphasize approaches for transcribing polyphonic music produced by pitched instruments and voice. Outside the scope of the article are methods for transcribing nonpitched sounds, such as drums, for which a brief overview is given in the "Percussion and Unpitched Sounds" section, as well as methods for transcribing specific sources within a polyphonic mixture, such as melody and bass lines.

Applications and impact

A successful AMT system would enable a broad range of interactions between people and music, including music education (e.g., through systems for automatic instrument tutoring), music creation (e.g., dictating improvised musical ideas and automatic music accompaniment), music production (e.g., music content visualization and intelligent content-based editing), music search (e.g., indexing and recommendation of music by melody, bass, rhythm, or chord progression), and musicology (e.g., analyzing jazz improvisations and other nonnotated music). As such, AMT is an enabling technology with clear potential for both economic and societal impact.

AMT is closely related to other music signal processing tasks [3], such as audio source separation, which also involves the estimation and inference of source signals from mixture observations. It is also useful for many high-level tasks in music information retrieval [4], such as structural segmentation, cover-song detection, and assessment of music similarity, since these tasks are much easier to address once the musical notes are known. Thus, AMT provides the main link between the fields of music signal processing and symbolic music processing (i.e., the processing of music notation and music language modeling). The integration of the two aforementioned fields through AMT will be discussed in the section "Further Extensions and Future Work."

Given the potential impact of AMT, the problem has attracted commercial interest in addition to academic research. While it is outside the scope of this article to provide a comprehensive list of commercial AMT software, commonly used applications include Melodyne (http://www.celemony.com/en/melodyne), AudioScore (http://www.sibelius.com/products/audioscore/), ScoreCloud (http://scorecloud.com/), AnthemScore (https://www.lunaverus.com/), and Transcribe! (https://www.seventhstring.com/xscribe/). It is worth noting that AMT papers in the literature have refrained from making explicit comparisons with commercially available music transcription software, possibly because of the difference in scope and target application between commercial and academic tools.

AMT is an enabling technology with clear potential for both economic and societal impact.

Analogies to other fields

AMT has close relations with other signal processing problems. With respect to the field of speech processing, AMT is widely considered to be the musical equivalent of automatic speech recognition (ASR), in the sense that both tasks involve converting acoustic signals to symbolic sequences. Like the cocktail party problem in speech, music usually involves multiple simultaneous voices, but, unlike speech, these voices are highly correlated in time and in frequency (see challenges 2 and 3 in the "Key Challenges" section). In addition, both AMT and ASR systems benefit from language modeling components that are combined with acoustic components to produce plausible results. Thus, there are also clear links between AMT and the wider field of natural language processing (NLP), with music having its own grammatical rules or statistical regularities, in a way similar to natural language [5]. The use of language models for AMT is detailed in the section "Further Extensions and Future Work."

Within the emerging field of sound scene analysis, there is a direct analogy between AMT and sound event detection (SED) [6], in particular with polyphonic SED, which involves detecting and classifying multiple overlapping events from audio. While everyday and natural sounds do not exhibit the same degree of temporal regularity and intersource frequency dependence as found in music signals, there are close interactions

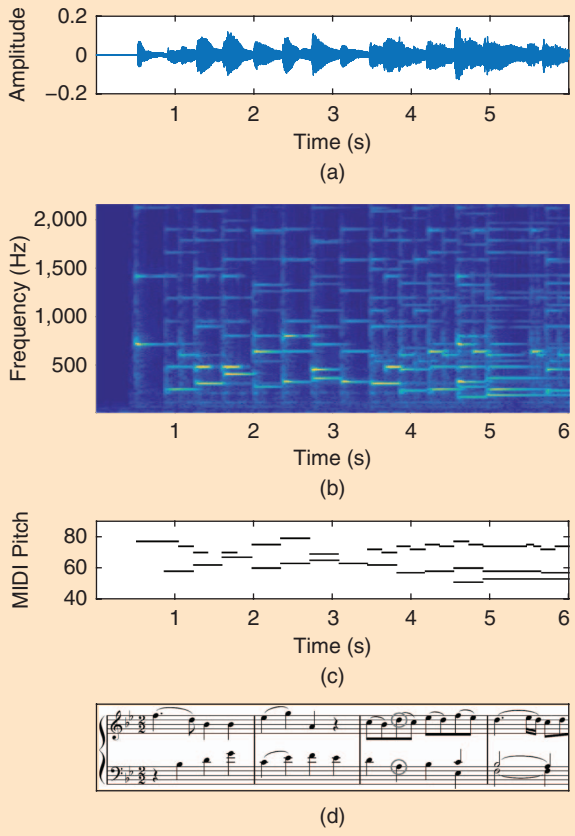

FIGURE 1. The data represented in an AMT system: the (a) input waveform, (b) internal time–frequency representation, (c) output piano-roll representation, and (d) output music score, with notes A and D marked in gray circles. The example corresponds to the first 6 s of W.A. Mozart's Piano Sonata no. 13, third movement. (Images courtesy of the MIDI Aligned Piano Sounds database.) MIDI: Musical Instrument Digital Interface.

between the two problems in terms of the methodologies used, as observed in the literature [6].

Furthermore, AMT is related to image processing and computer vision, as musical objects, such as notes, can be recognized as two-dimensional patterns in time–frequency representations. Compared with image processing and computer vision, where occlusion is a common issue, AMT systems are often affected by musical objects occupying the same time–frequency regions (this is detailed in the "Key Challenges" section).

Key challenges

Compared to other problems in the music signal processing field or the wider signal processing discipline, there are several factors that make AMT particularly challenging:

1) Polyphonic music contains a mixture of multiple simultaneous sources (e.g., instruments and vocals) with different pitch, loudness, and timbre (sound quality), with each source producing one or more musical voices. Inferring musical attributes (e.g., pitch) from the mixture signal is an extremely underdetermined problem.
2) Overlapping sound events often exhibit harmonic relations with each other. For any consonant musical interval, the fundamental frequencies form small integer ratios, so that their harmonics overlap in frequency, making the separation of the voices even more difficult. Taking a C major chord as an example, the fundamental frequency ratio of its three notes C:E:G is 4:5:6, and the percentages of harmonic positions that are overlapped by the other notes are 46.7%, 33.3%, and 60% for C, E, and G, respectively.
3) The timing of musical voices is governed by the regular metrical structure of the music. In particular, musicians pay close attention to the synchronization of onsets and offsets between different voices, which violates the common assumption of statistical independence between sources, which otherwise facilitates separation.
4) The annotation of ground-truth transcriptions for polyphonic music is very time consuming and requires high expertise. The lack of such annotations has limited the use of powerful supervised-learning techniques to specific AMT subproblems, such as piano transcription, where the annotation can be automated because of certain piano models that can automatically capture performance data. An approach to circumvent this problem was proposed in [7], but it requires professional music performers and thorough score pre- and postprocessing. We note that sheet music does not generally provide good ground-truth annotations for AMT; it is not time-aligned to the audio signal, nor does it usually provide an accurate representation of a performance. Even when accurate transcriptions exist, it is not trivial to identify corresponding pairs of audio files and musical scores because of the multitude of versions of any given musical work that are available from music distributors. At best, musical scores can be viewed as weak labels.

AMT provides the main link between the fields of music signal processing and symbolic music processing.

These key challenges are often not fully addressed in current AMT systems, leading to common issues in the AMT outputs, such as octave errors, semitone errors, missed notes (in particular, in the presence of dense chords), extra notes (often manifested as harmonic errors in the presence of unseen timbres), merged or fragmented notes, incorrect onsets/offsets, or misassigned streams [1], [2]. The remainder of this article will focus on ways to address the previously mentioned challenges as well as on discussion of additional open problems for the creation of robust AMT systems.

An overview of AMT methods

In the past four decades, many approaches have been developed for AMT for polyphonic music. While the end goal of AMT is to convert an acoustic music recording to some form of music notation, most approaches were designed to achieve a certain intermediate goal. Depending on the level of abstraction and the structures that need to be modeled for achieving such goals, AMT approaches can be generally organized into four categories: frame level, note level, stream level, and notation level.

Frame-level transcription, or MPE, is the estimation of the number and pitch of notes that are simultaneously present in each time frame (on the order of 10 ms). This is usually performed independently in each frame, although contextual information is sometimes considered through filtering frame-level pitch estimates in a postprocessing stage. Figure 2(a) shows an example of a frame-level transcription, where each black dot is a pitch estimate. Methods in this category do not form the concept of musical notes and rarely model any high-level musical structures.

A large portion of existing AMT techniques operate at this level. Recent approaches include traditional signal processing methods [11], [12], probabilistic modeling [8], Bayesian approaches [13], NMF [14]–[17], and neural networks (NNs) [18], [19]. All of these methods have pros and cons, and the research has not converged on a single approach. For example, traditional signal processing methods are simple and fast and generalize better to different instruments, while deep NN methods generally achieve higher accuracy on specific instruments (e.g., piano). Bayesian approaches provide a comprehensive modeling of the sound generation process, but the models can be very complex and slow. Readers interested in a comparison of the performance of different approaches are referred to the Multiple Fundamental Frequency Estimation and Tracking task of the annual Music Information Retrieval Evaluation eXchange (MIREX) (http://www.music-ir.org/mirex). However, readers are reminded that evaluation results may be biased by the limitations of data sets and evaluation metrics (see the sections "Key Challenges" and "Evaluation Metrics").

Note-level transcription, or note tracking, is one level higher than MPE in terms of the richness of structures of the estimates. It not only estimates the pitches in each time frame but also connects pitch estimates over time into notes. In the AMT literature,

a musical note is usually characterized by three elements: pitch, onset time, and offset time [1]. As note offsets can be ambiguous, they are sometimes neglected in the evaluation of note-tracking approaches, and, as such, some note-tracking approaches only estimate pitch and onset times of notes. Figure 2(b) shows an example of a note-level transcription, where each note is shown as a red circle (onset) followed by a black line (pitch contour). Many note-tracking approaches form notes by postprocessing MPE outputs (i.e., pitch estimates in individual frames). Techniques that have been used in this context include median filtering [12], hidden Markov models (HMMs) [20], and NNs [5]. This postprocessing is frequently performed for each Musical Instrument Digital Interface (MIDI) pitch independently without considering the interactions among simultaneous notes. This often leads to spurious or missing notes that share harmonics with correctly estimated notes.

From frame level to note level to stream level, the transcription task becomes more complex as more musical structures and cues need to be modeled.

Some approaches have been proposed to consider note interactions through a spectral likelihood model [9] or a music language model [5], [18] (see the "MLMs" section). Another subset of approaches estimates notes directly from the audio signal instead of building upon MPE outputs. Some approaches first detect onsets and then estimate pitches within each interonset interval [21], while others estimate pitch, onset, and sometimes offset in the same framework [22]–[24].

Stream-level transcription, also called *multipitch streaming* (*MPS*), targets the grouping of estimated pitches or notes into streams, where each stream typically corresponds to one instrument or musical voice and is closely related to instrument source separation. Figure 2(c) shows an example of a stream-level transcription, where pitch streams of different instruments have different colors. Compared to note-level transcription, the pitch contour of each stream is much longer than a single note and contains multiple discontinuities that are caused by silence, nonpitched sounds, and abrupt frequency changes. Therefore, techniques that are ordinarily used in note-level transcription are generally not sufficient for grouping pitches with a long and discontinuous contour. One important cue for MPS that is not explored in MPE and note tracking is timbre. Notes of the same stream (source) generally show similar timbral characteristics compared to those in different streams. Therefore, stream-level transcription is also called *timbre tracking* or *instrument tracking* in the literature. Existing works at this level are few, with [10], [16], and [25] as examples.

From frame level to note level to stream level, the transcription task becomes more complex as more musical structures and cues need to be modeled. However, the transcription outputs at these three levels are all parametric transcriptions, which are parametric descriptions of the audio content. The MIDI piano roll shown in Figure 1(c) is a good example of such a transcription. It is, indeed, an abstraction of music audio, but it has not yet reached the level of abstraction of music notation. Time is still measured in the unit of seconds instead of beats; pitch is measured in MIDI numbers instead of spelled note names that are compatible with the key (e.g., C♯ versus D♭); and the concepts of beat, bar, meter, key, harmony, and stream are lacking.

Notation-level transcription aims to transcribe the music audio into a human-readable musical score, such as the staff notation widely used in Western classical music. Transcription at this level requires a deeper understanding of musical structures, including harmonic, rhythmic, and stream structures. Harmonic structures, such as keys and chords, influence the note spelling of each MIDI pitch; rhythmic structures, such as beats and bars, help to quantize the length of notes; and stream structures aid the assignment of notes to different staffs. There has been some work on the estimation of musical structures from audio or MIDI

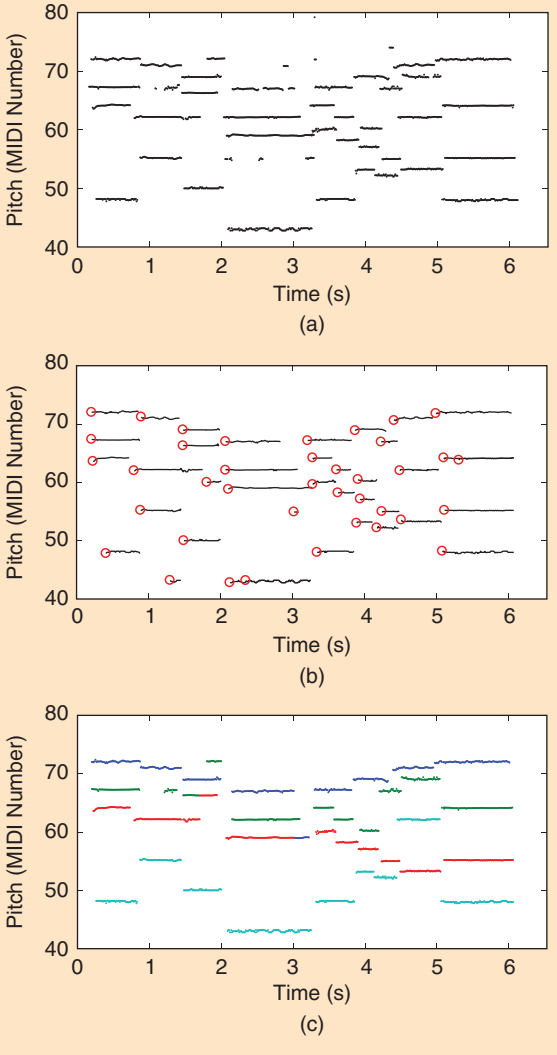

FIGURE 2. Examples of (a) frame-level, (b) note-level, and (c) stream-level transcriptions, produced by running the methods proposed in [8], [9], and [10], respectively, of the first phrase of J.S. Bach's chorale *Ach Gott und Herr* from the Bach10 data set. All three levels are parametric descriptions of the music performance.

representations of a performance. For example, methods for pitch spelling [26], timing quantization [27], and voice separation [28] from performed MIDI files have been proposed. However, little work has been done on integrating these structures into a complete music notation transcription, especially for polyphonic music.

Several software packages, including Finale, GarageBand, and MuseScore, provide the functionality of converting a MIDI file into music notation, but the results are typically not satisfying, and it is not clear what musical structures have been estimated and integrated during the transcription process. Cogliati et al. [29] proposed a method to convert a MIDI performance into music notation, based on a systematic comparison of the transcription performance with the aforementioned software. In terms of audio-to-notation transcription, a proof-of-concept work using end-to-end NNs was proposed by Carvalho and Smaragdis [30] to directly map music audio into music notation without explicitly modeling musical structures.

State of the art

While there is a wide range of applicable methods, AMT has been dominated during the last decade by two algorithmic families: NMF and NNs. Both families have been used for a variety of tasks, from speech and image processing to recommender systems and NLP. Despite this wide applicability, both approaches offer a range of properties that make them particularly suitable for modeling music recordings at the note level.

While there is a wide range of applicable methods, AMT has been dominated during the last decade by two algorithmic families: NMF and NNs.

NMF for AMT

The basic idea behind NMF and its variants is to represent a given nonnegative time–frequency representation $\mathbf{V} \in \mathbb{R}_{\geq 0}^{M \times N}$, e.g., a magnitude spectrogram, as a product of two nonnegative matrices: a dictionary $\mathbf{D} \in \mathbb{R}_{\geq 0}^{M \times K}$ and an activation matrix $\mathbf{A} \in \mathbb{R}_{\geq 0}^{K \times N}$ (see Figure 3). Computationally, the goal is to minimize a distance (or divergence) between $\mathbf{V}$ and $\mathbf{DA}$ with respect to $\mathbf{D}$ and $\mathbf{A}$. As a straightforward approach to solving this minimization problem, multiplicative update rules have been central to the success of NMF. For example, the generalized Kullback–Leibler divergence between $\mathbf{V}$ and $\mathbf{DA}$ is nonincreasing under the following updates and guarantees the nonnegativity of both $\mathbf{D}$ and $\mathbf{A}$ as long as both are initialized with positive real values [31]:

$$\mathbf{A} \leftarrow \mathbf{A} \odot \frac{\mathbf{D}^\top \left(\frac{\mathbf{V}}{\mathbf{DA}}\right)}{\mathbf{D}^\top \mathbf{J}} \quad \text{and} \quad \mathbf{D} \leftarrow \mathbf{D} \odot \frac{\left(\frac{\mathbf{V}}{\mathbf{DA}}\right)\mathbf{A}^\top}{\mathbf{J}\mathbf{A}^\top},$$

where the $\odot$ operator denotes pointwise multiplication, $\mathbf{J} \in \mathbb{R}^{M \times N}$ denotes the matrix of ones, and the division is pointwise. Intuitively, the update rules can be derived by choosing a

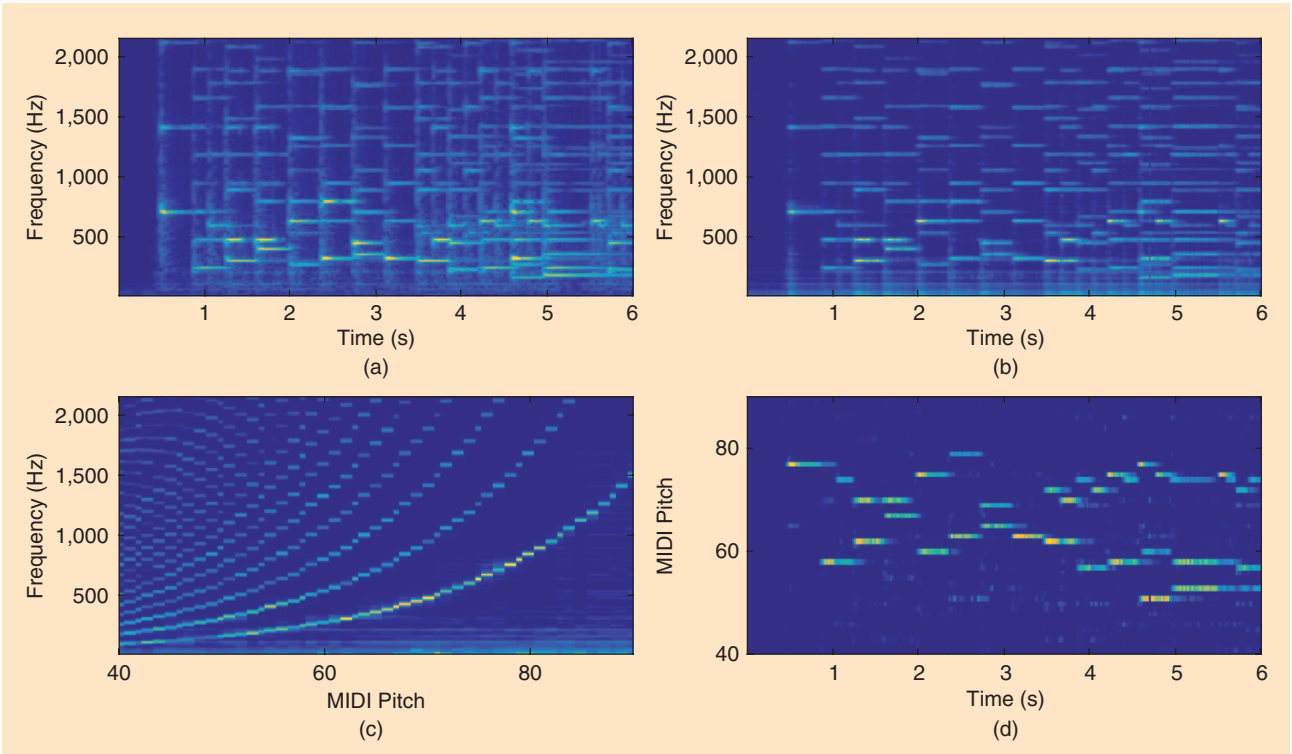

FIGURE 3. An example of NMF, using the same audio recording as in Figure 1: the (a) input spectrogram **V**, (b) approximated spectrogram **DA**, (c) dictionary **D** (preextracted), and (d) activation matrix **A**.

specific step size in a gradient (or, rather, coordinate) descent-based minimization of the divergence [31].

In an AMT context, both unknown matrices have an intuitive interpretation. The nth column of $\mathbf{V}$, i.e., the spectrum at time point n, is modeled in NMF as a linear combination of the K columns of $\mathbf{D}$, and the corresponding K coefficients are given by the nth column of $\mathbf{A}$. Given this point of view, each column of $\mathbf{D}$ is generally referred to as a *spectral template* and usually represents the expected spectral energy distribution associated with a specific note played on a specific instrument. For each template, the corresponding row in $\mathbf{A}$ is referred to as the associated *activation* and encodes when and how intensely that note is played over time. Given the nonnegativity constraints, NMF yields a purely constructive representation in the sense that the spectral energy modeled by one template cannot be canceled by another. This property is often seen as instrumental in identifying a parts-based and interpretable representation of the input [31].

In Figure 3, an NMF-based decomposition is illustrated. The magnitude spectrogram $\mathbf{V}$ shown in Figure 3(a) is modeled as a product of the dictionary $\mathbf{D}$ and activation matrix $\mathbf{A}$, shown in Figure 3(c) and (d), respectively. The product $\mathbf{DA}$ is given in Figure 3(b). In this case, the templates correspond to individual pitches, with clearly visible fundamental frequencies and harmonics. Additionally, comparing $\mathbf{A}$ with the piano-roll representation shown in Figure 1(c) indicates the correlation between NMF activations and the underlying musical score.

While Figure 3 illustrates the principles behind NMF, it also indicates why AMT is difficult. Indeed, a regular NMF decomposition would rarely look as clean as in Figure 3. Compared to speech analysis, sound objects in music are highly correlated. For example, even in a simple piece as in Figure 1, most pairs of simultaneous notes are separated by musically consonant intervals, which acoustically means that many of their partials overlap [e.g., the A and D notes around 4 s, marked with gray circles in Figure 1(d), share a high number of partials]. In this case, it can be difficult to disentangle how much energy belongs to which note. The task is further complicated by the fact that the spectrotemporal properties of notes vary considerably between different pitches, playing styles, dynamics, and recording conditions. Furthermore, the stiffness property of the strings affects the travel speed of transverse waves based on their frequency. As a result, the partials of instruments like the piano are not found at perfect integer multiples of the fundamental frequency. Because of this property, called *inharmonicity*, the positions of partials differ between individual pianos (see Figure 4).

To address these challenges, the basic NMF model has been extended by encouraging additional structure in the dictionary and the activations. For example, an important principle is to enforce sparsity in $\mathbf{A}$ to obtain a solution dominated by activations that are few but substantial; the success of sparsity paved the way for a whole range of sparse-coding approaches, in which the dictionary size K can considerably exceed the input dimension M [32]. Other extensions focus on the dictionary design. In the case of supervised NMF, the dictionary is precomputed and fixed using additionally available training material. For example, given K recordings, each containing only a single note, the dictionary shown in Figure 3(b) was constructed by extracting one template from each recording. This way, the templates are guaranteed to be free of interference from other notes and also to have a clear interpretation. As another example, Figure 5 illustrates an extension in which each NMF template is represented as a linear combination of fixed narrow-band subtemplates [15], which enforces a harmonic structure for all NMF templates. This way, a dictionary can be adapted to the recording to be transcribed, while maintaining its clean, interpretable structure.

In shift-invariant dictionaries, a single template can be used to represent a range of different fundamental frequencies. In particular, using a logarithmic frequency axis, the distances between individual partials of a harmonic sound are fixed, and thus shifting a template in frequency allows the modeling of sounds of varying pitch. Sharing parameters between different pitches in this way has turned out to be effective for increasing model capacity (see, e.g., [16] and [17]). Furthermore, spectrotemporal dictionaries alleviate a specific weakness of NMF models. In NMF, it is difficult to express that notes often have a specific temporal evolution. For example, the beginning of a note (or attack phase) might have entirely different spectral properties than the central part (decay phase). Such relationships are modeled in spectrotemporal dictionaries using a Markov process, which governs the sequencing of templates across frames so that different subsets of templates can be used for the attack and decay parts, respectively [16], [23].

NNs for AMT

As for many tasks relating to pattern recognition, NNs have, in recent years, had a considerable impact on the problem of music transcription and on music signal processing in general. NNs are able to learn a nonlinear function (or a composition of functions) from input to output via an optimization algorithm, such as stochastic gradient descent [33]. Compared to other fields, including image processing, progress on NNs for music transcription has been slower, and we will discuss a few of the underlying reasons.

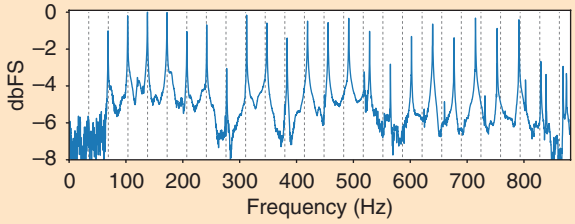

FIGURE 4. An illustration of inharmonicity: the spectrum of a C♯1 note played on a piano. The stiffness of the strings causes partials to be shifted from perfect integer multiples of the fundamental frequency (shown as vertical dotted lines). Here, the 23rd partial is at the position where the 24th harmonic would be expected. Note that the fundamental frequency of 34.65 Hz is missing, as piano soundboards typically do not resonate for modes with a frequency smaller than ≈ 50 Hz.

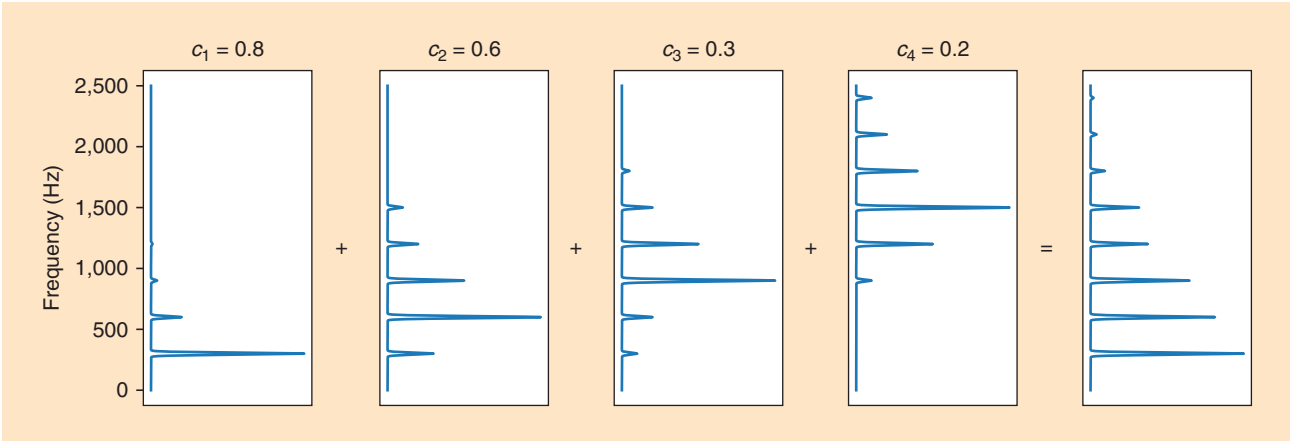

FIGURE 5. An illustration of harmonic NMF [15]. Each NMF template (far right) is represented as a linear combination of fixed narrow-band subtemplates. The resulting template is constrained to represent harmonic sounds by construction.

One of the earliest approaches based on NNs was Marolt's Sonic system [21]. A central component in this approach was the use of time-delay networks, which resemble convolutional networks in the time direction [33] and were employed to analyze the output of adaptive oscillators to track and group partials in the output of a gammatone filterbank. Although it was initially published in 2001, the approach remains competitive and still appears in comparisons in more recent publications [23].

In the context of the more recent revival of NNs, a first successful system was presented by Böck and Schedl [34]. One of the core ideas was to use two spectrograms as input to enable the network to exploit both a high time accuracy (when estimating the note onset position) and a high frequency resolution (when disentangling notes in the lower frequency range). This input is processed using one (or more) long short-term memory (LSTM) layers [33]. The potential benefit of using LSTM layers is twofold. First, the spectral properties of a note evolve across input frames, and LSTM networks have the capability to compactly model such sequences. Second, medium- and long-range dependencies between notes can potentially be captured. For example, based on a popular chord sequence, after hearing C and G major chords followed by A minor, a likely successor is an F major chord. An investigation of whether such long-range dependencies are indeed modeled, however, was not within the scope of this work.

Compared to other fields, including image processing, progress on NNs for music transcription has been slower.

Sigtia et al. [18] focused on long-range dependencies in music by combining an acoustic front end with a symbolic-level module resembling a language model, as used in speech processing. Using information obtained from MIDI files, a recurrent NN (RNN) was trained to predict the active notes in the next time frame, given those in the past. This approach needed to learn and represent a very large joint probability distribution, i.e., a probability for every possible combination of active and inactive notes across time. Note that, even in a single frame, there are 2^{88} possible combinations of notes on a piano. To render the problem of modeling such an enormous probability space tractable, the approach employed a specific NN architecture (the Neural Autoregressive Distribution Estimator, also known as *NADE*), which represented a large joint probability as a long product of conditional probabilities, an approach quite similar to the idea popularized recently by the well-known WaveNet architecture. Despite the use of a dedicated music language model, which was trained on relatively large MIDI-based data sets, only modest improvements over an HMM baseline could be observed, and thus the question remains open regarding to which degree long-range dependencies were indeed captured.

To further disentangle the influence of the acoustic front end from the language model on potential improvement in performance, Kelz et al. [19] focused on the acoustic modeling, reporting on the results of a larger-scale hyperparameter search and describing the influence of individual system components. Trained using this careful and extensive procedure, the resulting model outperformed existing models by a reasonable margin. In other words, while in speech processing, language models have led to a drastic improvement in performance, the same effect is still to be demonstrated in an AMT system, a challenge we will discuss in more detail hereafter.

The development of NN-based AMT approaches continues. The current state-of-the-art method for general-purpose piano transcription was proposed by Google Brain [24]. Combining and extending ideas from existing methods, this approach combines two networks (Figure 6). One detects note onsets, and its output is used to inform a second network, which focuses on perceiving note lengths. This can be interpreted from a probabilistic point of view. Note onsets are rare events compared to framewise note activity detections. The split into two network branches can thus be interpreted as splitting the representation of a relatively complex joint probability distribution over onsets and frame activity into a probability over onsets and a probability over frame activities, conditioned on the onset

distribution. Since the temporal dynamics of onsets and frame activities are quite different, this can lead to improved learning behavior for the entire network when trained jointly.

A comparison of NMF and NN models

Given the popularity of NMF and NN-based methods for AMT, it is interesting to discuss their differences. In particular, neglecting the nonnegativity constraints, NMF is a linear, generative model. Given that NMF-based methods are increasingly being replaced by NN-based ones, the question arises in which way linearity could be a limitation for an AMT model.

To look into this, assume we are given an NMF dictionary with two spectral templates for each musical pitch. To represent an observed spectrum of a single pitch C4, we can linearly combine the two templates associated with C4. The set (or manifold) of valid spectra for C4 notes, however, is complex, and thus, in most cases, our linear interpolation will not correspond to a real-world recording of a C4. We could increase the number of templates such that their interpolation could potentially get closer to a real C4—but the number of invalid spectra we can represent increases much more quickly compared to the number of valid spectra. Deep networks have shown considerable potential in recent years to implicitly represent such complex manifolds in a robust and comparatively efficient way [33]. An additional benefit over generative models like NMF is that NNs can be trained in an end-to-end fashion, i.e., note detections can be a direct output of a network without the need for additional postprocessing of model parameters (such as NMF activations).

Yet, despite these quite principled limitations, NMF-based methods remain competitive or even exceed the results achieved using NNs. Currently, there are two main challenges for NN-based approaches. First, there are only a few, relatively small annotated data sets available, and these are often subject to severe biases [7]. The largest publicly available data set [11] contains several hours of piano music—but all recorded on only seven different synthesizer-based and real pianos. While typical data augmentation strategies, such as pitch shifting or simulating different room acoustics, might mitigate some of the effects, there is still a considerable risk that a network overfits the acoustic properties of these specific instruments. For many types of instruments, even small data sets are not available. Other biases include musical style as well as the distribution over central musical concepts, such as key, harmony, tempo, and rhythm.

A second considerable challenge is the adaptability to new acoustic conditions. Providing just a few examples of isolated notes of the instrument to be transcribed, considerable improvements are observed in the performance of NMF-based models. There is currently no corresponding, equally effective mechanism to retrain or adapt an NN-based AMT system on a few seconds of audio. Thus, the error rate for nonadapted networks can be an order of magnitude higher than that of an adapted NMF system [23], [24]. Overall, as both of these challenges cannot easily be overcome, NMF-based methods are likely to remain relevant in specific use cases.

In Figure 7, we qualitatively illustrate some differences in the behavior of systems based on supervised NMF and NNs. Both systems were specifically trained for transcribing piano recordings, and we expose the approaches to a recording of an organ. Like the piano, the organ is played with a keyboard, but its acoustic properties are quite different. The harmonics of

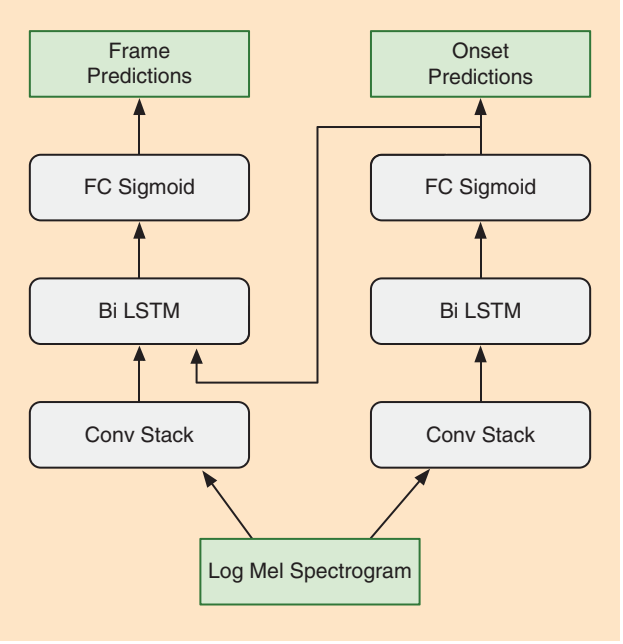

FIGURE 6. Google Brain's Onset and Frames Network. The input is processed by an initial network detecting note onsets. The result is used as side information for a second network focused on estimating note lengths. Bi LSTM: bidirectional LSTM layers; FC Sigmoid: fully connected sigmoid layer; Conv Stack: a series of convolutional layers. (Image adapted with permission from [24].)

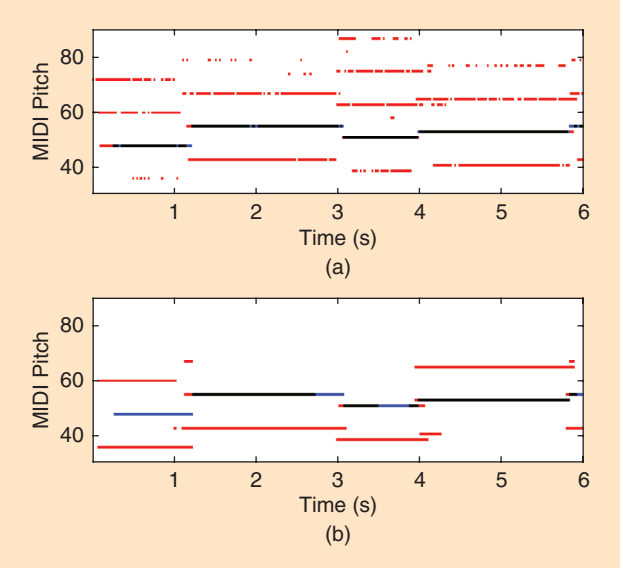

FIGURE 7. Piano-roll representations of the first 6 s of a recording of a Bach piece (BWV 582) for the organ. The black corresponds to correctly detected pitches, red to false positives, and blue to false negatives. (a) The output of an NMF model trained on piano templates. (b) The output of the piano-music-trained NN model of [24].

the organ are rich in energy and cover the entire spectrum; the energy of the notes does not decay over time, and the onsets are less pronounced. With this experiment, we want to find out how gracefully the systems fail when they encounter a sound that is outside the piano-sound manifold but still musically valid.

Comparing the NMF output in Figure 7(a) and the NN output in Figure 7(b) with the ground truth, we find that both methods detect additional notes (shown in red), mostly at octaves above and below the correct fundamental. Given the rich energy distribution, this behavior is expected. While we use a simple baseline model for NMF—and thus some errors could be attributed to that choice—the NN fails more gracefully. That is, fewer octave errors and fewer spurious short note detections are observed. (Yet, in terms of recall, the NMF-based approach identifies additional correct notes.)

It is difficult to argue why the acoustic model within the network should be better prepared for such a situation. However, the results suggest that the network learned something additional: the LSTM layers as used in the network (compare Figure 6) seem to have learned how typical piano notes evolve in time, and thus most note lengths look reasonable and less spurious. Similarly, the bandwidth in which octave errors occur is narrower for the NN, which could potentially indicate that the network models the likelihood of co-occurring notes or, in other words, a simple music language model (MLM). This leads us to our discussion of important remaining challenges in AMT.

As might be evident by surveying the AMT literature, the vast majority of approaches target only Western (or Eurogenetic) music.

Further extensions and future work

MLMs

As outlined in the "Analogies to Other Fields" section, AMT is closely related to ASR. In the same way that a typical ASR system consists of an acoustic component and a language component, an AMT system can model both the acoustic sequences and the underlying sequence of notes and other music cues over time. AMT systems have thus incorporated MLMs for modeling sequences of notes in a polyphonic context, with the aim of improving transcription performance. The capabilities of deep-learning methods toward modeling high-dimensional sequences have recently made polyphonic music sequence prediction possible. Boulanger-Lewandowski et al. [5] combined a restricted Bolzmann machine (RBM) with an RNN for polyphonic music prediction, which was used to postprocess the acoustic output of an AMT system.

Sigtia et al. [18] also used the aforementioned RNN–RBM as an MLM and combined the acoustic and language predictions using a probabilistic graphical model. While these initial works showed promising results, there are several directions for future research in MLMs. These include creating unified acoustic and language models (as opposed to using MLMs as postprocessing steps) and modeling other musical cues, such as chords, key, and meter (as opposed to simply modeling note sequences).

Score-informed transcription

If a known piece is performed, the musical score provides a strong prior for the transcription. In many cases, there are discrepancies between the score and a given music performance, which may be due to a specific interpretation by a performer or to performance mistakes. For applications like music education, it is useful to identify such discrepancies, by incorporating the musical score as additional prior information to simplify the transcription process (score-informed music transcription [35]). Typically, systems for such transcription use a score-to-audio alignment method as a preprocessing step to align the music score with the input music audio prior to performing transcription, as in [35]. While specific instances of such systems have been developed for certain instruments (piano and violin), the problem is still relatively unexplored, as is the related and more challenging problem of lead-sheet-informed transcription and the eventual integration of these methods toward the development of automatic music tutoring systems.

Context-specific transcription

While the creation of a blind multi-instrument AMT system without specific knowledge of the music style, instruments, and recording conditions is yet to be achieved, considerable progress has been reported on the problem of context-specific transcription, where prior knowledge of the sound of the specific instrument model or manufacturer and the recording environment is available. For context-specific piano transcription, multipitch detection accuracy can exceed 90% [22], [23], making such systems appropriate for user-facing applications. Open work in this topic includes the creation of context-specific AMT systems for multiple instruments.

Non-Western music

As might be evident by surveying the AMT literature, the vast majority of approaches target only Western (or Eurogenetic) music. This allows several assumptions, regarding both the instruments used and also the way that music is represented and produced. Typical assumptions include octaves containing 12 equally spaced pitches; two modes, major and minor; and a standard tuning frequency of A4 = 440 Hz.

However, these assumptions do not hold true for other music styles from around the world, where an octave is often divided into microtones (e.g., Arabic music theory is based on quartertones) or where there are modes not used in Western music (e.g., classical Indian music recognizes hundreds of modes, called *ragas*). Therefore, automatically transcribing non-Western music still remains an open problem with several challenges, including the design of appropriate signal and music notation representations while avoiding a so-called Western bias [36]. Another major issue is the lack of annotated data sets for non-Western music,

rendering the application of data-intensive machine-learning methods difficult.

Expressive pitch and timing

Western notation conceptualizes music as sequences of unchanging pitches being maintained for regular durations and has little scope for representing expressive use of microtonality and microtiming or for the detailed recording of timbre and dynamics. Research on automatic transcription has followed this narrow view, describing notes in terms of discrete pitches plus onset and offset times. For example, no suitable notation exists for performed singing, the most universal form of music making. Likewise, for other instruments without fixed pitch or with other expressive techniques, better representations are required. These richer representations can then be reduced to Western score notation, if required, by modeling musical knowledge and stylistic conventions.

Automatically transcribing non-Western music still remains an open problem with several challenges.

Percussion and unpitched sounds

An active problem in the music signal processing literature is that of detecting and classifying nonpitched sounds in music signals [1, Ch. 5]. In most cases, this is expressed as the problem of drum transcription, since the vast majority of contemporary music contains mixtures of pitched sounds and unpitched sounds produced by a drum set. Drum set components typically include the bass drum, snare drum, hi-hat, cymbals, and toms. The problem in this case is to detect and classify percussive sounds into one of the aforementioned sound classes. Elements of the drum transcription problem that make it particularly challenging are the concurrent presence of several harmonic, inharmonic, and nonharmonic sounds in the music signal, as well as the requirement of an increased temporal resolution for drum transcription systems compared to typical multipitch detection systems. Approaches for pitched instrument transcription and drum transcription have largely been developed independently, and the creation of a robust music transcription system that supports both pitched and unpitched sounds still remains an open problem.

Evaluation metrics

Most AMT approaches are evaluated using the set of metrics proposed for the MIREX Multiple-F0 Estimation and Note Tracking public evaluation tasks (http://www.music-ir.org/mirex/). Three types of metrics are included: frame based, note based, and stream based, mirroring the frame-level, note-level, and stream-level transcription categories presented in the "State of the Art" section. While the above sets of metrics all have their merits, it could be argued that they do not correspond with human perception of music transcription accuracy, where, e.g., an extra note might be considered as a more severe error than a missed note, or where out-of-key note errors might be penalized more compared with in-key ones. Therefore, the creation of perceptually relevant evaluation metrics for AMT and the creation of evaluation metrics for notation-level transcription remain open problems.

Conclusions

AMT has remained an active area of research in the fields of music signal processing and music information retrieval for several decades, with several potential benefits in other areas and fields extending beyond music. As outlined in this article, there remain several hurdles to be overcome, i.e., on modeling music signals and on the availability of data, as described in the "Key Challenges" section; with respect to the limitations of state-of-the-art methodologies, as described in the section "A Comparison of NMF and NN-Models"; and, finally, on extensions beyond the current area of existing tasks, as presented in the "Further Extensions and Future Work" section. We believe that addressing these challenges will lead toward the creation of a complete music transcription system and unlock the full potential of music signal processing technologies. Supplementary audio material related to this article can be found on the companion website (http://c4dm.eecs.qmul.ac.uk/spm-amt-overview/).

Acknowledgment

Emmanouil Benetos is supported by U.K. RAEng Research Fellowship RF/128. The authors are listed alphabetically.

Authors

Emmanouil Benetos (emmanouil.benetos@qmul.ac.uk) received his B.Sc. and M.Sc. degrees in informatics from Aristotle University of Thessaloniki, Greece, in 2005 and 2007, respectively, and his Ph.D. degree in electronic engineering from Queen Mary University of London, in 2012. He is a lecturer and Royal Academy of Engineering research fellow with the Centre for Digital Music, Queen Mary University of London, and a Turing Fellow with the Alan Turing Institute. From 2013 to 2015, he was a university research fellow with the Department of Computer Science, City University of London. He has published more than 80 peer-reviewed papers spanning several topics in audio and music signal processing. His research focuses on signal processing and machine learning for music and audio analysis as well as applications to music information retrieval, acoustic scene analysis, and computational musicology. He is a Member of the IEEE.

Simon Dixon (s.e.dixon@qmul.ac.uk) received his B.Sc. (with honors) and Ph.D. degrees in computer science from the University of Sydney, Australia, in 1989 and 1994, respectively. He studied classical guitar at the New South Wales Conservatorium of Music, Sydney, where he obtained his A.Mus.A and L.Mus.A degrees in 1987 and and 1988, respectively. He is the deputy director of the Centre for Digital Music, Queen Mary University of London, where he is also a professor. His research is in music informatics, including high-level music signal analysis, computational modeling of musical knowledge, and the study of musical performance. Particular areas of focus include automatic music transcription, beat tracking, audio alignment, and analysis of intonation and temperament. He was president (2014–2015) of the International Society for