Automatic Music Transcription: An Overview

Emmanouil Benetos *Member, IEEE*, Simon Dixon, Zhiyao Duan *Member, IEEE*, and Sebastian Ewert *Member, IEEE*

I. Introduction

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 (multi-)pitch estimation, 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 (MIR) [1], [2]. Due to the very nature of music signals, which often contain several sound sources (e.g., musical instruments, voice) that produce one or more concurrent sound events (e.g., notes, 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 simultaneous notes[1] and multiple instruments [2].

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

In this paper, we provide a high-level overview of Automatic Music Transcription, 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 non-negative matrix factorization. Finally we provide an extensive discussion of open challenges for AMT. Regarding the scope of the paper, we emphasize approaches for transcribing polyphonic music produced by pitched instruments and voice. Outside the scope of the paper are methods for transcribing non-pitched sounds such as drums, for which a brief overview is given in Section IV-F, as well as methods for transcribing specific sources within a polyphonic mixture such as melody and bass line.

A. Applications & 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 non-notated 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 estimation and inference of source signals from mixture observations. It is also useful for many high-level tasks in MIR [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., processing of music notation and music language modeling). The integration of the two aforementioned fields through AMT will be discussed in Section IV.

Given the potential impact of AMT, the problem has also attracted commercial interest in addition to academic research. While it is outside the scope of the paper to provide a comprehensive list of commercial AMT software, commonly used software includes Melodyne[2], AudioScore[3], ScoreCloud[4], AnthemScore[5], and Transcribe![6]. It is worth noting that AMT papers in the literature have refrained from making explicit comparisons with commercially available music transcription software, possibly due to different scopes and target applications between commercial and academic tools.

B. 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

Authors in alphabetical order.
EB and SD are with the Centre for Digital Music, Queen Mary University of London, UK. e-mail: {emmanouil.benetos,s.e.dixon}@qmul.ac.uk.
ZD is with the Department of Electrical and Computer Engineering, University of Rochester, NY, USA. e-mail: zhiyao.duan@rochester.edu
SE is with Spotify Ltd, UK. e-mail: sewert@spotify.com
EB is supported by a UK RAEng Research Fellowship (RF/128).
[1]Called *polyphonic music* in the music signal processing literature.

[2]http://www.celemony.com/en/melodyne
[3]http://www.sibelius.com/products/audioscore/
[4]http://scorecloud.com/
[5]https://www.lunaverus.com/
[6]https://www.seventhstring.com/xscribe/

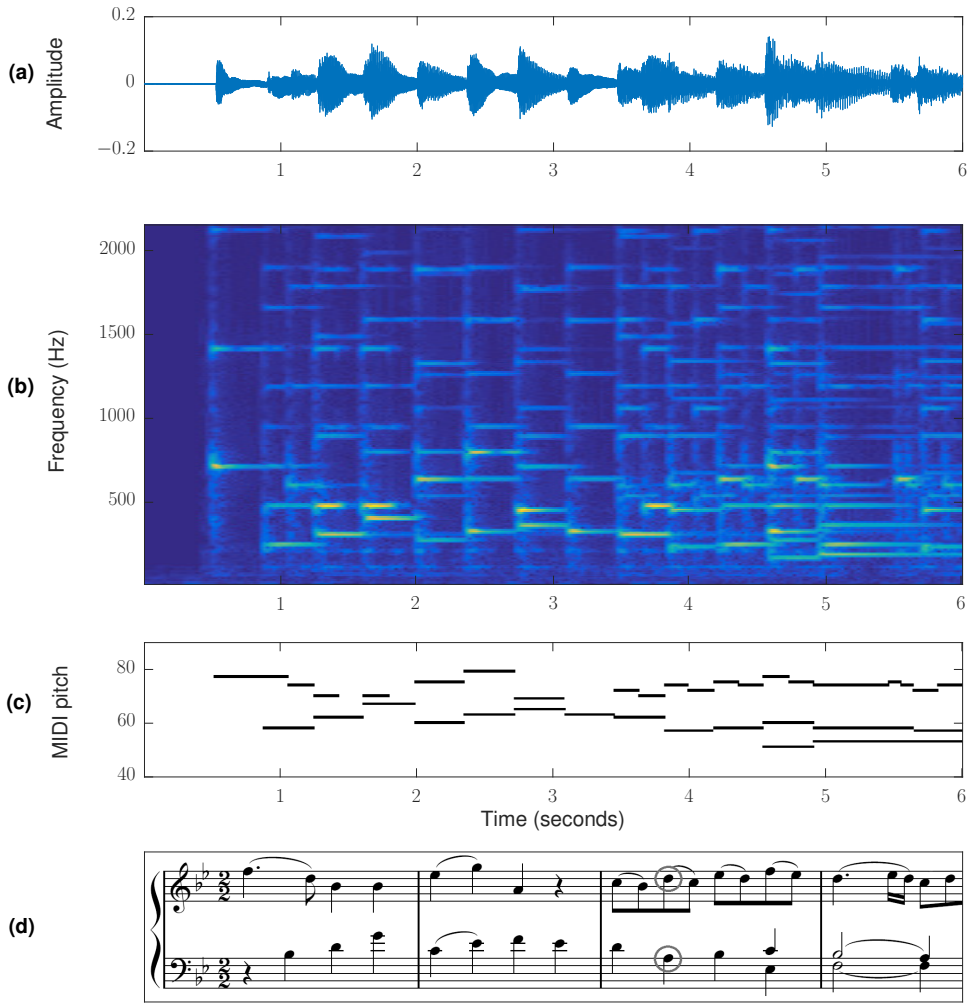


Figure 1. Data represented in an AMT system. (a) Input waveform, (b) Internal time-frequency representation, (c) Output piano-roll representation, (d) Output music score, with notes A and D marked in gray circles. The example corresponds to the first 6 seconds of W. A. Mozart's Piano Sonata No. 13, 3rd movement (taken from the MAPS database).

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 Section I-C). In addition, both AMT and ASR systems benefit from language modeling components that are combined with acoustic components in order 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 similar way to natural language [5]. The use of language models for AMT is detailed in Section IV.

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 inter-source frequency dependence as found in music signals, there are close interactions between the two problems in terms of the methodologies used, as observed in the literature [6].

Further, 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 Section I-C).

C. 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, 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 under-determined 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 percentage 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 sub-problems such as piano transcription, where the annotation can be automated due to certain piano models that can automatically capture performance data. An approach to circumvent this problem was proposed in [7], however, it requires professional music performers and thorough score pre- and post-processing. 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.

The above 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 mis-assigned streams [1], [2]. The remainder of the paper will focus on ways to address the above challenges, as well as discussion of additional open problems for the creation of robust AMT systems.

II. 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 *Multi-Pitch Estimation (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 in each frame independently, although contextual information is sometimes considered through filtering frame-level pitch estimates in a post-processing stage. Fig. 2(top) 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 approaches operate at this level. Recent approaches include traditional signal processing methods [11], [12], probabilistic modeling [8], Bayesian approaches [13], non-negative matrix factorization (NMF) [14], [15], [16], [17], and neural networks [18], [19]. All of these methods have pros and cons and the research has not converged to a single approach. For example, traditional signal processing methods are simple and fast and generalize better to different instruments, while deep neural network methods generally achieve higher accuracy on specific instruments (e.g., piano). Bayesian approaches provide a comprehensive modeling of the sound generation process, however 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 & Tracking task of the annual Music Information Retrieval Evaluation eXchange (MIREX) [7]. However, readers are reminded that evaluation

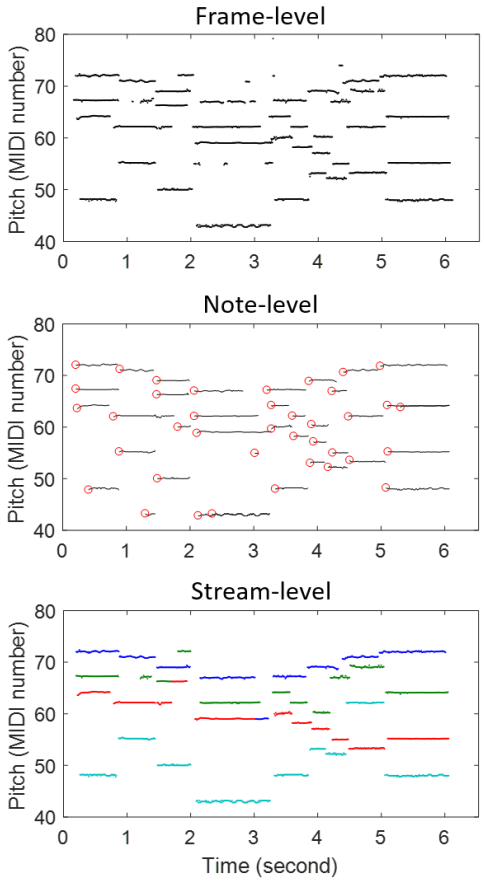


Figure 2. Examples of frame-level, note-level and stream-level transcriptions, produced by running methods proposed in [8], [9] and [10], respectively, of the first phrase of J. S. Bach's chorale "Ach Gott und Herr" (taken from the Bach10 dataset). All three levels are parametric descriptions of the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

[7] http://www.music-ir.org/mirex

results may be biased by the limitations of datasets and evaluation metrics (see Sections I-C and IV-G).

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 often 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. Fig. 2(middle) 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 post-processing 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 neural networks [5]. This post-processing is often performed for each 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. Some approaches have been proposed to consider note interactions through a spectral likelihood model [9] or a music language model [5], [18] (see Section IV-A). Another subset of approaches estimate notes directly from the audio signal instead of building upon MPE outputs. Some approaches first detect onsets and then estimate pitches within each inter-onset interval [21], while others estimate pitch, onset and sometimes offset in the same framework [22], [23], [24].

Stream-level transcription, also called *Multi-Pitch Streaming (MPS)*, targets grouping 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. Fig. 2(bottom) 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, non-pitched sounds and abrupt frequency changes. Therefore, techniques that are often used in note-level transcription are generally not sufficient to group pitches into 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 [16], [10], [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 Fig. 1(c) is a good example of such a transcription. It is indeed an abstraction of music audio, however, 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♯ vs 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 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 lengths 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 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, however, the results are often 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, with a systematic comparison of the transcription performance with the above-mentioned software. In terms of audio-to-notation transcription, a proof-of-concept work using end-to-end neural networks was proposed by Carvalho and Smaragdis [30] to directly map music audio into music notation without explicitly modeling musical structures.

III. State-of-the-art

While there is a wide range of applicable methods, automatic music transcription has been dominated during the last decade by two algorithmic families: Non-Negative Matrix Factorization (NMF) and Neural Networks (NNs). Both families have been used for a variety of tasks, from speech and image processing to recommender systems and natural language processing. Despite this wide applicability, both approaches offer a range of properties that make them particularly suitable for modeling music recordings at the note level.

A. Non-negative Matrix Factorization for AMT

The basic idea behind NMF and its variants is to represent a given non-negative time-frequency representation $\mathbf{V} \in \mathbb{R}_{\geq 0}^{M \times N}$, e.g., a magnitude spectrogram, as a product of two non-negative 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 Fig. 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 non-increasing under the following updates and

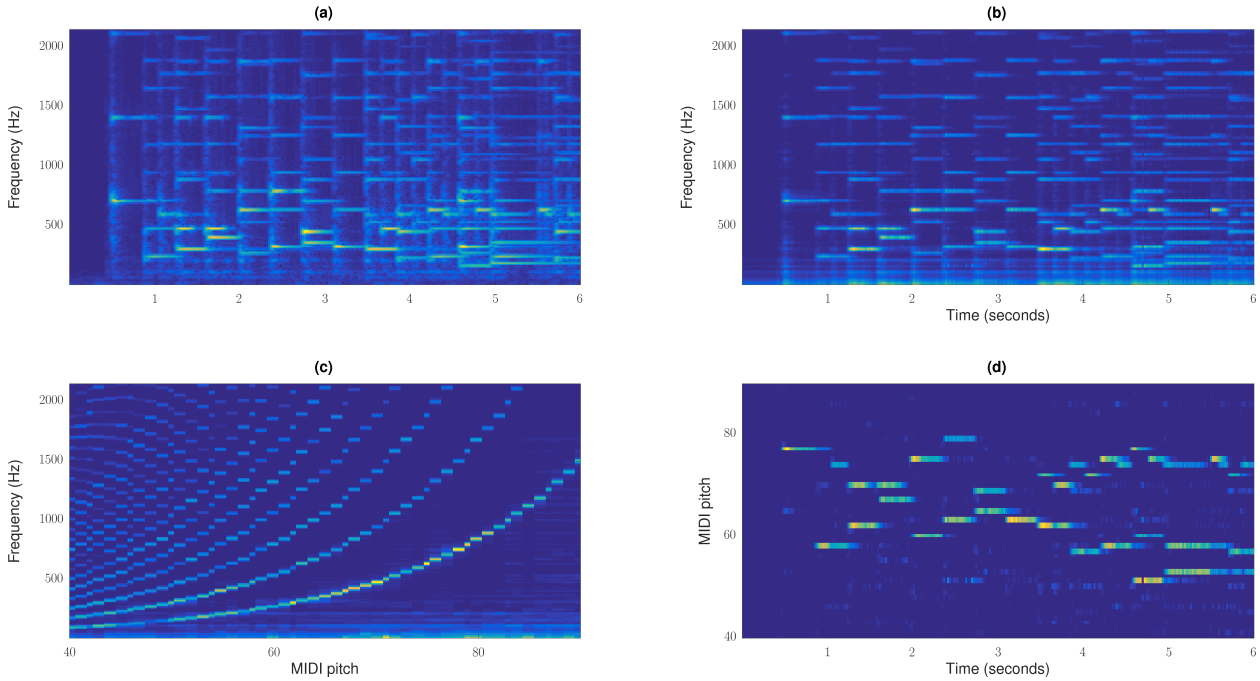


Figure 3. NMF example, using the same audio recording as Fig. 1. (a) Input spectrogram $\mathbf{V}$, (b) Approximated spectrogram $\mathbf{DA}$, (c) Dictionary $\mathbf{D}$ (pre-extracted), (d) Activation matrix $\mathbf{A}$.

guarantees the non-negativity of both $\mathbf{D}$ and $\mathbf{A}$ as long as both are initialized with positive real values [31]:

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

where the $\odot$ operator denotes point-wise multiplication, $\mathbf{J} \in \mathbb{R}^{M\times N}$ denotes the matrix of ones, and the division is point-wise. Intuitively, the update rules can be derived by choosing 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 n-th 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 n-th column of $\mathbf{A}$. Given this point of view, each column of $\mathbf{D}$ is often 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 non-negativity constraints, NMF yields a purely constructive representation in the sense that spectral energy modeled by one template cannot be cancelled by another – this property is often seen as instrumental in identifying a parts-based and interpretable representation of the input [31].

In Fig. 3, an NMF-based decomposition is illustrated. The magnitude spectrogram $\mathbf{V}$ shown in Fig. 3(a) is modeled as a product of the dictionary $\mathbf{D}$ and activation matrix $\mathbf{A}$ shown in Fig. 3(c) and (d), respectively. The product $\mathbf{DA}$ is given in Fig. 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 Fig. 1(c) indicates the correlation between NMF activations and the underlying musical score.

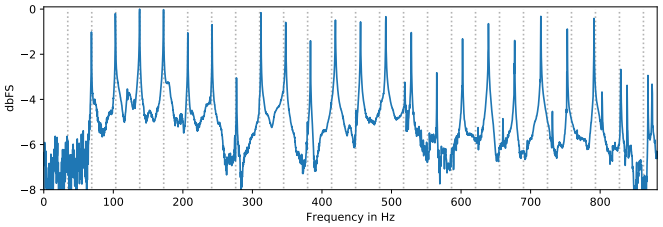


Figure 4. Inharmonicity: Spectrum of a C♯1 note played on a piano. The stiffness of strings causes partials to be shifted from perfect integer multiples of the fundamental frequency (shown as vertical dotted lines); here the 23rd partial is at the position where the 24th harmonic would be expected. Note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

While Fig. 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 Fig. 3. Compared to speech analysis, sound objects in music are highly correlated. For example, even in a simple piece as shown in Fig. 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 seconds, marked with gray circles in Fig. 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 spectro-temporal properties of notes vary considerably between different pitches, playing styles, dynamics and recording conditions. Further, stiffness properties of strings affect the travel speed of transverse waves based on their frequency – as a result,

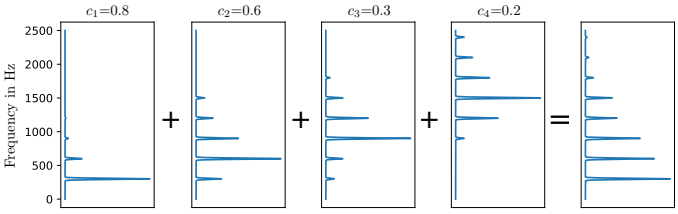


Figure 5. Harmonic NMF [15]: Each NMF template (right hand side) is represented as a linear combination of fixed narrow-band sub-templates. The resulting template is constrained to represent harmonic sounds by construction.

the partials of instruments such as the piano are not found at perfect integer multiples of the fundamental frequency. Due to this property called *inharmonicity*, the positions of partials differ between individual pianos (see Fig. 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 few but substantial activations – the success of sparsity paved the way for a whole range of *sparse coding* approaches, in which the dictionary size K can exceed the input dimension M considerably [32]. Other extensions focus on the dictionary design. In the case of supervised NMF, the dictionary is pre-computed and fixed using additionally available training material. For example, given K recordings each containing only a single note, the dictionary shown in Fig. 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 have a clear interpretation. As another example, Fig. 5 illustrates an extension in which each NMF template is represented as a linear combination of fixed narrow-band sub-templates [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 modeling sounds of varying pitch. Sharing parameters between different pitches in this way has turned out to be effective towards increasing model capacity (see e.g., [16], [17]). Further, *spectro-temporal* dictionaries alleviate a specific weakness of NMF models: in NMF it is difficult to express that notes often have a specific temporal evolution – e.g., 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 spectro-temporal 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 the decay parts, respectively [16], [23].

B. Neural Networks for AMT

As for many tasks relating to pattern recognition, neural networks (NNs) have had a considerable impact in recent years on the problem of music transcription and on music signal processing in general. NNs are able to learn a non-linear 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 below.

One of the earliest approaches based on neural networks was Marolt's Sonic system [21]. A central component in this approach was the use of time-delay (TD) networks, which resemble convolutional networks in the time direction [33], and were employed to analyse the output of adaptive oscillators, in order 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 neural networks, 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 two-fold. 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 in scope.

Sigtia et al. [18] focus 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 network is trained to predict the active notes in the next time frame given the past. This approach needs 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 employs a specific neural network architecture (NADE), which represents a large joint 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 datasets, only modest improvements over an HMM baseline could be observed and thus the question remains open to which degree long-range dependencies are 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] focus on the acoustic modeling and report on the results of a larger scale hyperparameter search

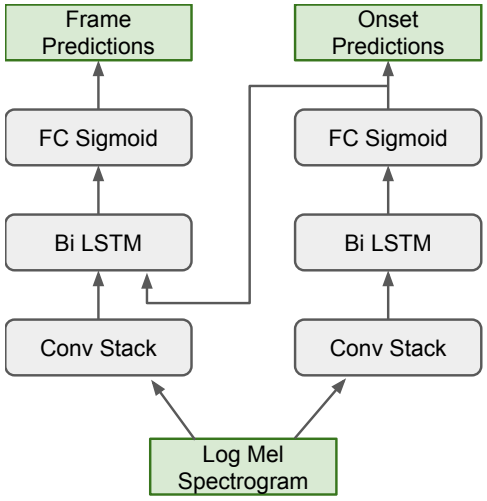


Figure 6. Google Brain's Onset and Frames Network: The input is processed by a first network detecting note onsets. The result is used as side information for a second network focused on estimating note lengths (adapted from [24]). *Bi LSTM* refers to bi-directional LSTM layers; *FC Sigmoid* refers to a fully connected sigmoid layer; *Conv Stack* refers to a series of convolutional layers.

and describe the influence of individual system components. Trained using this careful and extensive procedure the resulting model outperforms 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 below.

The development of neural network 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 (Fig. 6): one network is used to detect note onsets and its output is used to inform a second network, which focuses on detecting note lengths. This can be interpreted from a probabilistic point of view: note onsets are rare events compared to frame-wise 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.

C. A Comparison of NMF and Neural Network Models

Given the popularity of NMF and neural network based methods for automatic music transcription, it is interesting to discuss their differences. In particular, neglecting the non-negativity constraints, NMF is a linear, generative model. Given that NMF-based methods are increasingly 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 – however, 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 such as NMF is that neural networks 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 post-processing of model parameters (such as NMF activations).

Yet, despite these quite principled limitations, NMF-based methods remain competitive or even exceed results achieved using neural networks. Currently, there are two main challenges for neural network-based approaches. First, there are only few, relatively small annotated datasets available, and these are often subject to severe biases [7]. The largest publicly available dataset [11] contains several hours of piano music – however, 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 datasets 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 re-train or adapt a neural network based AMT system on a few seconds of audio – thus the error rate for non-adapted 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 Fig. 7, we qualitatively illustrate some differences in the behavior of systems based on supervised NMF and neural networks. 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 the organ are rich in energy and cover the entire spectrum, the energy of notes does not decay over time and 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 Fig. 7(a) and the NN output in Fig. 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.

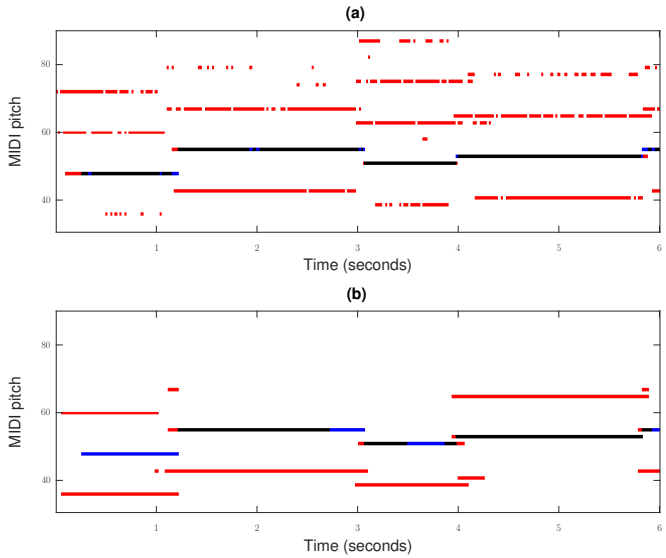


Figure 7. Piano-roll representations of the first 6 seconds of a recording of a Bach piece (BWV 582) for organ. Black color corresponds to correctly detected pitches, red to false positives, and blue to false negatives. (a) Output of NMF model trained on piano templates. (b) Output of the piano music-trained neural network model of [24].

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 neural network fails more gracefully: 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 to such a situation. However, the results suggest that the network learned something additional: the LSTM layers as used in the network (compare Fig. 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 neural network, which could potentially indicate that the network models the likelihood of co-occurring notes or, in other words, a simple music language model, which leads us to our discussion of important remaining challenges in AMT.

IV. Further Extensions and Future Work

A. Music Language Models

As outlined in Section I-B, AMT is closely related to automatic speech recognition (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 also the underlying sequence of notes and other music cues over time. AMT systems have thus incorporated *music language models (MLMs)* for modeling sequences of notes in a polyphonic context, with the aim of improving transcription performance. The capabilities of deep learning methods towards 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 post-process 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 post-processing steps) and modeling other musical cues, such as chords, key and meter (as opposed to simply modeling note sequences).

B. 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 due to performance mistakes. For applications such as 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 score-informed music transcription use a *score-to-audio alignment* method as a pre-processing step, in order to align the music score with the input music audio prior to performing transcription, e.g. [35]. While specific instances of score-informed transcription systems have been developed for certain instruments (piano, 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 towards the development of automatic music tutoring systems.

C. 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, multi-pitch detection accuracy can exceed 90% [23], [22], making such systems appropriate for user-facing applications. Open work in this topic includes the creation of context-specific AMT systems for multiple instruments.

D. 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; a standard tuning frequency of A4 = 440 Hz). However, these assumptions do not hold true for other music styles from around the world, where for instance an octave is often divided into *microtones* (e.g., Arabic music theory is based on quartertones), or on the existence of modes that are 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 datasets for non-Western music, rendering the application of data-intensive machine learning methods difficult.

E. 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, nor for 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.

F. Percussion and Unpitched Sounds

An active problem in the music signal processing literature is that of detecting and classifying non-pitched 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 kit. Drum kit 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 non-harmonic sounds in the music signal, as well as the requirement of an increased temporal resolution for drum transcription systems compared to typical multi-pitch 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.

G. 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[8]. 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 Sec. III. 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, as well as the creation of evaluation metrics for notation-level transcription, remain open problems.

V. Conclusions

Automatic music transcription 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 the remit of music. As outlined in this paper, there remain several challenges to be addressed in order to fully address this problem: these include key challenges as described in Section I-C on modeling music signals and on the availability of data, challenges with respect to the limitations of state-of-the-art methodologies as described in Section III-C, and finally on extensions beyond the current remit of existing tasks as presented in Section IV. We believe that addressing these challenges will lead towards the creation of a "complete" music transcription system and towards unlocking the full potential of music signal processing technologies. Supplementary audio material related to this paper can be found in the companion website[9].

References

[1] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription*. New York: Springer, 2006.

[2] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, Dec. 2013.

[3] M. Müller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, Oct. 2011.

[4] M. Schedl, E. Gómez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Foundations and Trends in Information Retrieval*, vol. 8, pp. 127–261, 2014.

[5] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[6] T. Virtanen, M. D. Plumbley, and D. P. W. Ellis, Eds., *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

[7] L. Su and Y.-H. Yang, "Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription," in *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2015, pp. 309–321.

[8] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, 2010.

[9] Z. Duan and D. Temperley, "Note-level music transcription by maximum likelihood sampling." in *ISMIR*, 2014, pp. 181–186.

[10] Z. Duan, J. Han, and B. Pardo, "Multi-pitch streaming of harmonic sound mixtures," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 138–150, Jan 2014.

[11] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[12] L. Su and Y.-H. Yang, "Combining spectral and temporal representations for multipitch estimation of polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1600–1612, Oct 2015.

[13] P. H. Peeling, A. T. Cemgil, and S. J. Godsill, "Generative spectrogram factorization models for polyphonic piano transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 519–527, March 2010.

[8] http://www.music-ir.org/mirex/

[9] http://c4dm.eecs.qmul.ac.uk/spm-amt-overview/

[14] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[15] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[16] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a temporally-constrained shift-invariant model," *Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1727–1741, March 2013.

[17] B. Fuentes, R. Badeau, and G. Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1854–1866, Sept 2013.

[18] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 927–939, May 2016.

[19] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2016, pp. 475–481.

[20] J. Nam, J. Ngiam, H. Lee, and M. Slaney, "A classification-based polyphonic piano transcription approach using learned feature representations," in *ISMIR*, 2011, pp. 175–180.

[21] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

[22] A. Cogliati, Z. Duan, and B. Wohlberg, "Context-dependent piano music transcription with convolutional sparse coding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2218–2230, Dec 2016.

[23] S. Ewert and M. B. Sandler, "Piano transcription in the studio using an extensible alternating directions framework," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 1983–1997, Nov 2016.

[24] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2018.

[25] V. Arora and L. Behera, "Multiple F0 estimation and source clustering of polyphonic music audio using PLCA and HMRFs," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 2, pp. 278–287, 2015.

[26] E. Cambouropoulos, "Pitch spelling: A computational model," *Music Perception*, vol. 20, no. 4, pp. 411–429, 2003.

[27] H. Grohganz, M. Clausen, and M. Mueller, "Estimating musical time information from performed MIDI files," in *Proceedings of International Society for Music Information Retrieval Conference*, 2014.

[28] I. Karydis, A. Nanopoulos, A. Papadopoulos, E. Cambouropoulos, and Y. Manolopoulos, "Horizontal and vertical integration/segregation in auditory streaming: a voice separation algorithm for symbolic musical data," in *Proceedings of Sound and Music Computing Conference (SMC)*, 2007.

[29] A. Cogliati, D. Temperley, and Z. Duan, "Transcribing human piano performances into music notation," in *Proceedings of the International Society for Music Information Retrieval Conference*, 2016, pp. 758–764.

[30] R. G. C. Carvalho and P. Smaragdis, "Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2017, pp. 151–155.

[31] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems (NIPS)*, 2001, pp. 556–562.

[32] S. A. Abdallah and M. D. Plumbley, "Unsupervised analysis of polyphonic music by sparse coding," *IEEE Transactions on neural Networks*, vol. 17, no. 1, pp. 179–196, 2006.

[33] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. MIT press Cambridge, 2016.

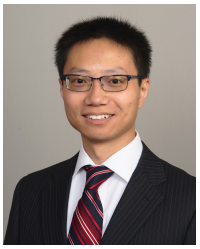
[34] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*, 2012, pp. 121–124.

[35] S. Wang, S. Ewert, and S. Dixon, "Identifying missing and extra notes in piano recordings using score-informed dictionary learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1877–1889, Oct 2017.

[36] X. Serra, "A multicultural approach in music information research," in *12th International Society for Music Information Retrieval Conference*, 2011, pp. 151–156.

Emmanouil Benetos (S'09, M'12) is Lecturer and Royal Academy of Engineering Research Fellow with the Centre for Digital Music, Queen Mary University of London, and Turing Fellow with the Alan Turing Institute. He received the Ph.D. degree in Electronic Engineering from Queen Mary University of London, U.K., in 2012. From 2013 to 2015, he was University Research Fellow with the Department of Computer Science, City, University of London. He has published over 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.

Simon Dixon is Professor and Deputy Director of the Centre for Digital Music at Queen Mary University of London. He has a Ph.D. in Computer Science (Sydney) and L.Mus.A. diploma in Classical Guitar. 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 (2014-15) of the International Society for Music Information Retrieval (ISMIR), is founding Editor of the Transactions of ISMIR, and has published over 160 refereed papers in the area of music informatics.



Zhiyao Duan (S'09, M'13) is an assistant professor in the Electrical and Computer Engineering Department at the University of Rochester. He received his B.S. in Automation and M.S. in Control Science and Engineering from Tsinghua University, China, in 2004 and 2008, respectively, and received his Ph.D. in Computer Science from Northwestern University in 2013. His research interest is in the broad area of computer audition, i.e., designing computational systems that are capable of understanding sounds, including music, speech, and environmental sounds. He co-presented a tutorial on Automatic Music Transcription at ISMIR 2015. He received a best paper award at the 2017 Sound and Music Computing (SMC) conference and a best paper nomination at the 2017 International Society for Music Information Retrieval (ISMIR) conference.

Sebastian Ewert is a Senior Research Scientist at Spotify. He received the M.Sc./Diplom and Ph.D. degrees (summa cum laude) in computer science from the University of Bonn (svd. Max-Planck-Institute for Informatics), Germany, in 2007 and 2012, respectively. In 2012, he was awarded a GAES fellowship and joined the Centre for Digital Music, Queen Mary University of London (United Kingdom). At the Centre, he became Lecturer for Signal Processing in 2015 and was one of the founding members of the Machine Listening Lab, which focuses on the development of machine learning and signal processing methods for audio and music applications.