From Audio to Music Notation

24

Lele Liu and Emmanouil Benetos

24.1 Introduction

The field of Music Information Retrieval (MIR) focuses on creating methods and practices for making sense of music data from various modalities, including audio, video, images, scores and metadata [54]. Within MIR, a core problem which to the day remains open is Automatic Music Transcription (AMT), the process of automatically converting an acoustic music signal into some form of musical notation. The creation of a method for automatically converting musical audio to notation has several uses including but also going beyond MIR: from software for automatic typesetting of audio into staff notation or other music representations, to the use of automatic transcriptions as a descriptor towards the development of systems for music recommendation, to applications for interactive music systems such as automatic music accompaniment, for music education through methods for automatic instrument tutoring, and towards enabling musicological research in sound archives, to name but a few.

Interest in AMT has grown during recent years as part of recent advances in artificial intelligence and in particular deep learning, which have led to new applications, systems, as well as have led to a new set of technical, methodological and ethical challenges related to this problem. This chapter presents state-of-the-art research and open topics in AMT, focusing on recent methods for addressing this task based on deep learning, as well as on outlining challenges and directions for future research.

The first attempts to address this problem come back to the 1970s and the dawn of the field of computer music (e.g. [47]), while the problem faced a resurgence in the

L. Liu (✉) · E. Benetos
School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK
e-mail: lele.liu@qmul.ac.uk

E. R. Miranda (ed.), *Handbook of Artificial Intelligence for Music*,
https://doi.org/10.1007/978-3-030-72116-9_24

mid-2000s with the development of methods for audio signal processing and pattern recognition, and encountered a second wave of popularity in recent years following the emergence of deep learning methods. Irrespective of the methodologies used to investigate and develop tools and practices for AMT, researchers addressing this task draw knowledge from several disciplines, including digital signal processing, machine learning/artificial intelligence, music perception and cognition, musical acoustics and music theory. There are also strong links with other problems both within and beyond MIR, including Optical Music Recognition (OMR—which is the counterpart of AMT but for printed music or manuscripts instead of recorded audio—e.g. [50]), automatic speech recognition and speaker diarisation [66], sound event detection for everyday and nature sounds [59], and object recognition and tracking in video [17]. AMT is also closely related to the fields of music language modelling and symbolic music processing [15], serving as a bridge between the acoustic and symbolic domains in music.

Given the complexity of the problem of AMT, the overarching task is often split into subtasks, including pitch/multi-pitch detection, onset and offset detection, instrument identification and tracking, meter estimation and rhythm quantisation, estimation of dynamics and expression and typesetting/engraving. However, recent advances in artificial intelligence have promoted the development of 'end-to-end' methods for AMT, thus often skipping intermediate tasks or steps and directly producing a transcription in a particular notation format. Figure 24.1 shows the typical stages of an AMT system for a short excerpt from a Mozart sonata, starting with the input waveform, the extracted time-frequency representation (in this case a short-time Fourier transform magnitude spectrogram), the output transcription in piano-roll representation and the output transcription in the form of Western staff notation.

Despite active research on this problem for decades and measurable progress over the years, AMT is still faced by several challenges, both technical and ethical. Broadly, the performance of certain AMT systems can be deemed sufficient for audio recordings containing solo acoustic instruments, within the context of Western tonal music, assuming a relatively moderate tempo and a level of polyphony around 3, 4. Here, the term 'polyphony' refers to the maximum number of concurrent pitches at a given time instant. The problem of automatically transcribing audio recordings which contain sounds produced by multiple instruments, vocals and percussion with a high polyphony level or a fast tempo is still relatively limited. Other factors that can affect the performance of such systems include the existence of distortions either at the instrumental production stage or at the audio production/mastering stage, or cases where the performance or composition in question does not fall under the auspices of Western tonal music. A relatively new challenge which has emerged with the adoption of data-driven methods for addressing the task is the bias imposed by the algorithms through the choice of datasets. Given that most datasets for AMT include Western tonal music performed by solo piano or other solo Western orchestral instruments have created certain limits and biases with respect to the range of instruments or to the range of music cultures and styles that can be supported by state-of-the-art AMT systems. Limitations of symbolic representations and encodings for music (MIDI, MEI, MusicXML, Lilypond, etc.) also further constrain the potential of current AI-

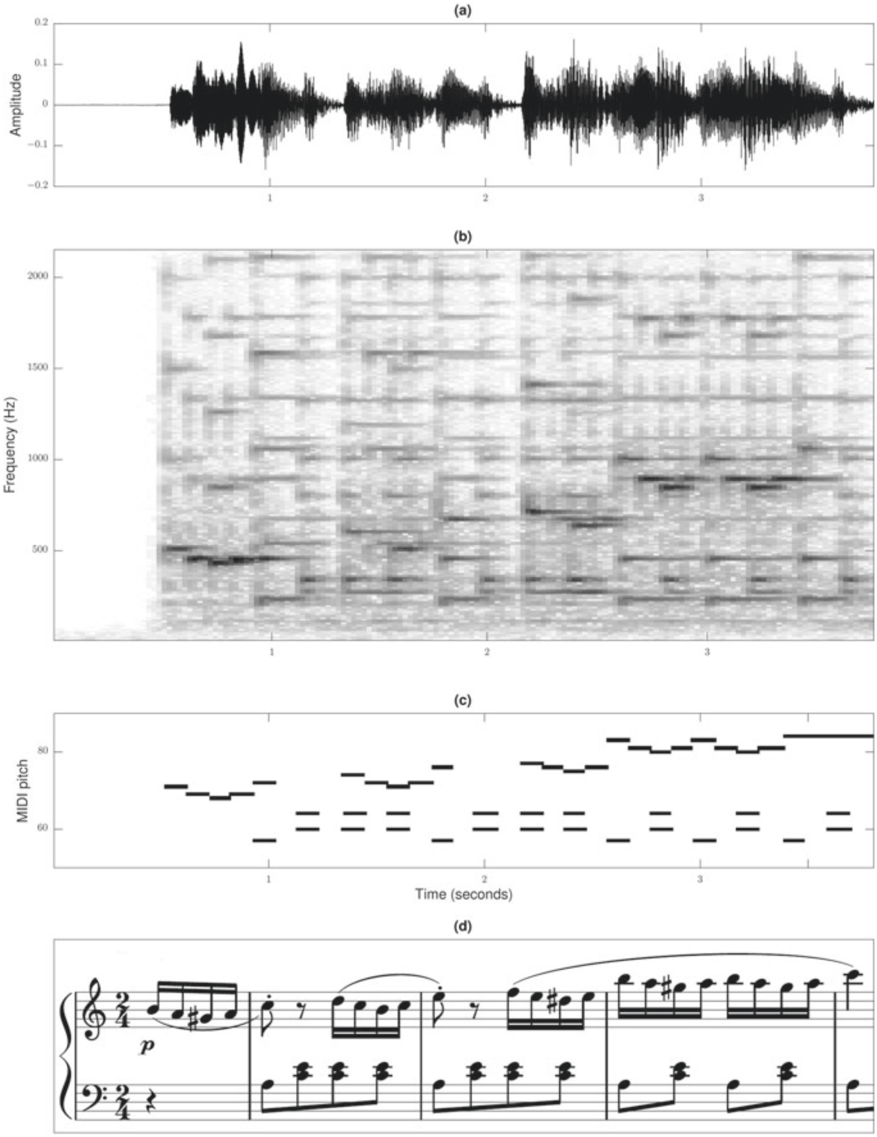

Fig. 24.1 Typical stages of an AMT system: **a** input waveform; **b** time-frequency representation; **c** output piano-roll representation; **d** output music score, typeset using Musescore. The example corresponds to the first 4 s of W.A. Mozart's Piano Sonata no. 11, 3rd movement

based AMT systems to support the transcription of music performances that cannot necessarily be expressed through Western staff notation or do not assume 12-tone equal temperament.

The aim of this chapter is to provide a review and discussion of recent methods for AMT, focusing on methods based on AI and deep learning in particular. The focus of the chapter is on automatic transcription of pitched sounds; see [61] for a recent review on the related task of Automatic Drum Transcription (ADT). For a detailed look on signal processing and statistical methods for AMT, the reader is referred to [36]; for a discussion related to the challenges of AMT methods relying on signal processing or statistical methods, please see [6]. A recent tutorial-like overview of both 'traditional' machine learning and deep learning methodologies for AMT is presented in [5].

The outline of this chapter is as follows. Section 24.2 provides a concise definition of various problems that have been posed under AMT; an overview of commonly used datasets and evaluation metrics in AMT is presented in Sect. 24.3. An overview of the state-of-the-art in AMT is presented in Sect. 24.4, including a more detailed look at deep learning methods for the task. Current methodological and ethical challenges facing AMT methods, tools, systems and practices are outlined in Sect. 24.5. Finally, conclusions are presented in Sect. 24.6.

24.2 Problem Definition

As mentioned in Sect. 24.1, AMT is divided into several subtasks, and most approaches have only been addressing a small subset of these subtasks. Perhaps the most essential subtask (especially when referring to the transcription of pitched sounds) is *pitch detection*, or in the case of multiple concurrent sounds, *multi-pitch detection*. Here, we define pitch in the same way as in [27], where a sound has a certain pitch if it can be reliably matched to a sine tone of a given frequency at a sound pressure level of 40 dB. Typically, this task refers to estimating one or more pitches at each time frame (e.g. at 10 ms intervals), where pitch is typically expressed in Hz. Given the close links between pitch and the fundamental frequency of periodic signals, this task is often referred to as multiple-F0 estimation. This task is publicly evaluated annually as part of the Music Information Retrieval Evaluation eXchange (MIREX) task on MultiF0 estimation [1].

It is often useful for multi-pitch detection systems to produce a non-binary representation of estimated pitches over time, which could be used for pitch visualisation purposes, or as an intermediate feature for other MIR tasks that rely on an initial pitch estimate (e.g. melody estimation [53], chord estimation [41]). Often this representation is referred to as *pitch salience*, or a *time-pitch representation*. Figure 24.2a shows the pitch salience representation for the excerpt of Fig. 24.1 using the method of [8].

Moving on to a higher level of abstraction which is closer to how humans might transcribe music, we would need express notes as characterised by their start time, end

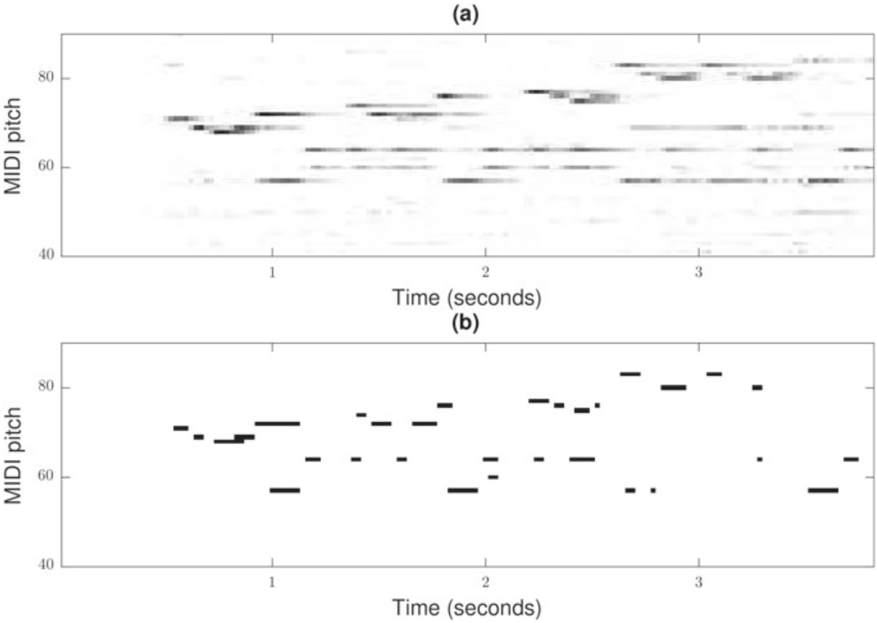

Fig. 24.2 **a** The pitch salience representation for the excerpt of Fig. 24.1 using the method of [8]; **b** The corresponding binarised piano-roll representation

time, and pitch—in a similar way as expressed, e.g. in the MIDI format. This task is referred to as *note tracking* and involves the subtasks of *onset detection* (i.e. detecting the start of a note), *offset detection* (i.e. detecting the end of a note), and (multi-)pitch detection. A comprehensive tutorial on signal processing-based methods for onset detection can be found in [4]. Approaches for note tracking are publicly evaluated annually as part of the Music Information Retrieval Evaluation eXchange (MIREX) note tracking task [1]. Figure 24.2b shows the output of the note tracking process by performing simple thresholding on the pitch salience of Fig. 24.2a.

In addition to the detection of pitched sounds and their timings, a key element towards a successful musical transcription is on assigning each detected note to the musical instrument that produced it. This task is referred to in the literature as *instrument assignment*, *timbre tracking*, or *multi-pitch streaming*. A closely related task in the wider field of MIR is that of *musical instrument recognition* from audio, which has received relatively little attention from the research community (see [30] for a recent overview).

The above mentioned note tracking task estimates the start and end times of notes, but in terms of seconds as opposed to beats or any other metrical subdivision. To that end, the task of *rhythm transcription* or *note value recognition* aims to estimate the metrical structure of the music recording in question and estimate the note timings and durations in terms of metrical subdivisions (e.g. [20,44]). By having estimated pitches with their respective timings in terms of meter, one can typeset the transcribed

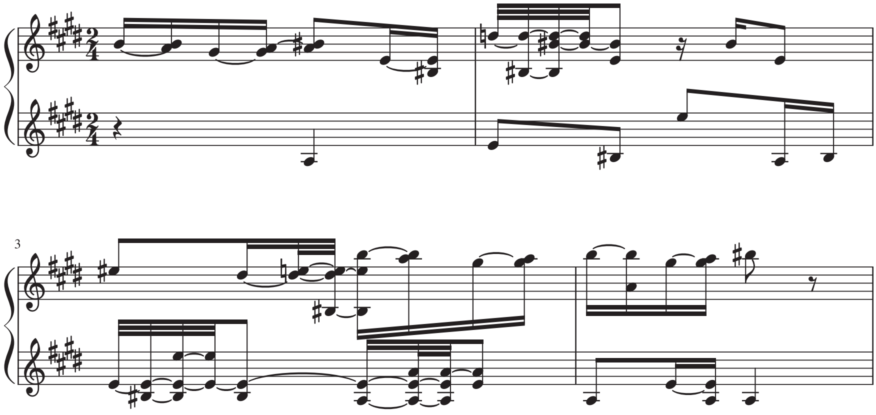

Fig. 24.3 The rhythm-quantised transcription of the excerpt of Fig. 24.1, automatically transcribed using the method of [8] and typeset using Musescore (https://musescore.org/)

audio in some form of human-readable music notation, e.g. Western staff notation. This is a task that depending on the complexity of the music performance in question might also require to split the detected stream of notes into multiple music staves (this is referred to as *voice separation* and *staff estimation*). The process of converting music audio into staff notation is sometimes referred to as *complete music transcription* (taking into account that such a 'complete' transcription might not contain information related to musical instruments, phrasing, expression or dynamics).

Figure 24.3 shows the rhythm-quantised transcription of the excerpt of Fig. 24.1 in Western staff notation, automatically transcribed using the method of [8]. While from a first glance there are little similarities with the score of Fig. 24.1d, a close inspection shows that the majority of pitches have been correctly detected, although their respective durations are not properly estimated (which can be attributed to sustain and pedalling of the piano performance of this piece).

24.3 Datasets and Evaluation Metrics

24.3.1 Datasets

As there are an increasing amount of exploration on deep learning methods for AMT, people are using larger datasets to train and evaluate the systems they developed. There are several datasets that are commonly used for AMT problems in literature, such as the RWC dataset [26], MIDI Aligned Piano Sounds (MAPS) dataset [24], Bach10 [22], MedleyDB [10], and MusicNet [58]. Two recently proposed datasets are MAESTRO [29] and Slakh [38]. Table 24.1 provides an overview on commonly used AMT datasets.

Table 24.1 AMT datasets and their properties. Instrument abbreviations—Vc: Vocal, Gt: Guitar, Bs: Bass, Dr: Drums, Pn: Piano, Tp: Trumpet, Cl: Cello, Vl: Violin, Cr: Clarinet Sx: Saxophone, Bn: Bassoon. Music style abbreviations—Cls: Classical, Plr: Popular, Jzz: Jazz, Ryf: Royalty-Free

Dataset	Instruments	Music style	Size	Comments
RWC dataset ([26], 2002)	Gt, Vc, Dr, Pn, Tp, Cl, etc.	Cls, Ryf, Plr, Jzz, etc.	315 pieces in 6 subsets	Real recordings, with non-aligned MIDI files for Popular, Royalty-Free, Classical and Jazz subsets. A version of automatically aligned MIDI annotations for the Classical subset can be found in the SyncRWC dataset [2]
MAPS dataset ([24], 2010)	Pn	Cls + non musical piece (notes and chords)	30 pieces * 9 piano synthesizers in the MUS subset	synthesized and real piano recordings. Additional rhythm and key annotations can be found in A-MAPS dataset [64]
Bach10 ([22], 2010)	Vl, Cr, Sx, Bn	Four-part J.S. Bach Chorales	10 pieces	Real recordings, individual stems, F0 annotations
MedleyDB ([10], 2014)	multiple instrument (Pn, Vc, etc.)	Ryf	196 pieces in MedleyDB 2.0	Real recordings, With individual stems of each instrument recording. 108 pieces with melody annotation
MusicNet ([58], 2016)	multiple instrument (Pn, Vl, Cl, etc.)	Cls	330 pieces	Real recordings under various conditions. Labels aligned by dynamic time wrapping and verified by trained musicians, estimated labeling error rate 4%
GuitarSet ([62], 2018)	Gt	Plr	360 pieces	Real recordings
MAESTRO ([29], 2019)	Pn	mostly Cls	1282 pieces	From e-piano competition, 201.2 h in total
Slakh ([38], 2019)	Pn, Gt, Bs, Dr, etc.	Cls, Plr, etc.	2100 tracks	Synthesized from Lakh MIDI dataset [49]

Although there are plenty of choices of AMT datasets, there are relatively more datasets for piano transcription (given the ease in automatically exporting MIDI annotations from acoustic pianos when using specific piano models such as Disklavier or Bösendorfer), but much less for other instruments, especially non-Western instruments. The biggest challenge of collecting AMT datasets is that annotating music recordings requires a high degree of music expertise, and is very time-consuming. Also, there might not be enough music pieces and recordings for some less popular traditional instruments when a large dataset is needed. Moreover, human-annotated transcription datasets are not guaranteed to have a high degree of temporal precision, which makes them less suitable for model evaluation on frame and note level. Su and Yang [57] proposed four aspects to evaluate the goodness of a dataset: generality, efficiency, cost and quality. They suggest that a good dataset should be not limited to a certain music form or recording conditions, should be fast-annotated, should be as low-cost as possible and be accurate enough. Because of the difficulty in collecting large human-transcribed datasets, researchers have used electronic instruments or acoustic instruments with sensors that can directly produce annotations (e.g. electronic piano, MAESTRO dataset), or synthesised datasets (e.g. Slakh) instead of real recordings. The use of synthesised recordings greatly speed up dataset collection, but on the other hand, could introduce some bias in model training, limiting generality of the developed AMT system.

24.3.2 Evaluation Metrics

Despite collecting datasets, model evaluation is another important process in developing methodologies for AMT problems. Evaluating a music transcription can be difficult since there are various types of errors, from pitch errors to missing/extra notes, and each has a different influence on the final evaluation of results. Currently, common evaluation metrics for AMT systems focus mainly on frame/note level transcriptions [3,9,14,28,33]. Much less work has been down on stream and notation level transcriptions [39,40,42]. In the 2019 annual Music Information Retrieval Evaluation eXchange (MIREX), there are three subtasks [1] for music transcription for pitched-instruments—multiple fundamental frequency estimation on frame level, note tracking and timbre tracking (multi-pitch streaming).

Common multiple fundamental frequency estimation methods [3] calculate frame-wise *precision*, *recall* and relevant *F-measure* values. The three scores are defined as:

$$precision = \frac{TP}{TP + FP} \tag{24.1}$$

$$recall = \frac{TP}{TP + FN} \tag{24.2}$$

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall} \tag{24.3}$$

The *TP*, *FP* and *FN* values correspond to *true positives*, *false positives* and *false negatives* respectively, and are calculated from all pitch values and time frames in the piano roll. There are also other methods for evaluating frame-wise transcription, such as separating different types of errors (e.g. missed pitches, extra pitches, false alarm) in multiple F0 estimation. A type-specific error rate is calculated in [48], where the authors defined a frame-level transcription error score combining different error types. Separating different error types can lead to a better interpretation on music transcription evaluation.

Note tracking problems usually define transcription results as sequences of notes, characterised by a pitch, onset and offset. A *tolerance* is defined to allow small errors in onset times since it is difficult to estimate exact time when building an AMT dataset as well as transcribing music with an AMT system. A common *tolerance* is 50 ms, which is used in the MIREX note tracking subtask. There are also some other scenarios where offset times are included (e.g. in [7] a 20% tolerance for offset is applied and in [19] a tolerance of the larger one in 20% of the note length or 50 ms is used for offset time). For any of the above scenarios, note-level precision, recall and F-measure are calculated for a final evaluation. Similar to frame-level F0 estimation, researchers have attempted to include error types in evaluation metrics (see e.g. [42]).

There are less works on *multi-pitch streaming*. The evaluation for *multi-pitch streaming* uses similar metric like precision and recall. Gómez and Bonada [25] proposed a simple method of calculating accuracy and false rate to evaluate voice streaming applied to A Capella transcription. In 2014, Duan and Temperley [23] used a similar evaluation method to calculate a more general multi-pitch streaming accuracy. The accuracy is defined as:

$$accuracy = \frac{TP}{TP + FP + FN} \tag{24.4}$$

Another work by Molina et al. [42] proposed to include types of errors in streaming process, and used a standard precision-recall metric.

Recent years has seen some introduction of evaluation metrics for *complete music transcription* given a recent increase in methods that directly transcribes audio to music scores. Some methods proposed include [18,39,40]. A recent approach for evaluating score transcriptions is proposed by Mcleod and Yoshii [40], which is based on a previous approach [39] called *MV2H* (representing Multi-pitch detection, Voice separation, Metrical alignment, note Value detection and Harmonic analysis). According to this metric, a score is calculated for each of the five aspects, then the scores are combined into a joint evaluation following a principle of one mistake should not be penalised more than once.

While most of evaluation metrics are based on music theory and simple statistical analysis, there are some metrics that contain some considerations on human perception of music transcriptions. In 2008, Daniel et al. [21] explored the difference of some error types in AMT from the aspect of human perception, and proposed a modified evaluation metric that weights different error types.

24.4 State of the Art

In this section, we look into state-of-the-art methodologies for AMT, mainly focusing on Neural Network methods. The section will be structured as follows. In Sect. 24.4.1, we provide an overview for the development and common methods for AMT, followed by Sect. 24.4.2 where we discuss Neural Network methods used in AMT. The following sections cover more specific topics within AMT: we give a review on *multi-task learning methods* for AMT in Sect. 24.4.3; the use of *music language models* and related works are covered in Sect. 24.4.4 and finally we review works on *complete transcription* in Sect. 24.4.5.

24.4.1 Overview

As the field of MIR has evolved over the past 20 years since the inception of the International Symposium on Music Information Retrieval (ISMIR), so has the topic of AMT. Roughly, proposed methods for AMT in the early 2000s made use of signal processing and statistical machine learning theory (see [36] for more details). Following the seminal paper of [56] on the potential of non-negative matrix factorisation when applied to the problem of AMT, a series of different methods were proposed for AMT that made use of matrix decomposition approaches. In the early 2010s, following the rise of deep learning methods and the paper by Humphrey et al. [30] advocating for the use of deep learning methods for MIR, neural network-based methods started being widely used for AMT and are still in use to date.

In terms of AMT subtasks to be addressed, the vast majority of methods have been and still do focus on (framewise) multi-pitch detection, with a smaller proportion of methods focusing on note tracking or rhythm transcription/typesetting. Due to the emergence of end-to-end deep learning methods for AMT, an increasing trend towards systems producing higher-level representations (such as outputs in MIDI format or in staff notation) can be observed [5]. The problem of timbre tracking/instrument assignment is however still under-explored.

Current literature for AMT includes a mixture of deep learning and matrix decomposition approaches, with deep learning methods currently being used in the majority of scenarios. Compared to other tasks in MIR, a large proportion of methods still employ matrix decomposition approaches (see e.g. [5]), due to their ability to work with limited data, fast learning and inference, and due to the models' interpretability. The remainder of this chapter will focus more on neural network-based methods for AMT, due to their increasing popularity in the research community and also due to certain methodological challenges when using deep learning methods for AMT that are still to be addressed.

24.4.2 Neural Networks for AMT

Research in AMT has increasingly been relying on deep learning models, which use feedforward, recurrent and convolutional layers as main architectural blocks. An early example of a deep neural model applied to AMT is the work of Nam et al. [45], which uses a Deep Belief Network (DBN) in order to learn representations for a polyphonic piano transcription task. Resulting learned features are then fed to a Support Vector Machine (SVM) classifier in order to produce a final decision. Another notable early work that made use of deep neural architectures was by Böck and Schedl [13], where the authors used a bi-directional Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) units, applied to the task of polyphonic piano transcription. Two points are particularly worth mentioning for the work of [13]: (i) the use of two STFT magnitude spectrograms with different window sizes as inputs to the network, in order to achieve both a 'good temporal precision and a sufficient frequency resolution'; (ii) The output is a piano-roll representation of note onsets and corresponding pitches, and does not include information on note durations/offsets.

A first systematic study towards the use of various neural network architectures for AMT was done by Sigtia et al. in [55]. The study compared networks for polyphonic piano transcription that used feedforward, recurrent and convolutional layers (noting that layer types were not combined), all using a Constant-Q Transform (CQT) spectrogram as input time-frequency representation. Results from [55] showed that networks that include convolutional layers reported the best results for the task, which is also in line with other results reported in the literature, and with current methodological trends related to neural networks for AMT. The ability of Convolutional Neural Networks (CNNs) to function well for tasks related to multi-pitch detection and AMT stems from the useful property of shift-invariance in log-frequency representations such as the CQT: a convolutional kernel that is shifted across the log-frequency axis can capture spectro-temporal patterns that are common across multiple pitches.

Following the work of [55], Kelz et al. [33] showed the potential of simple frame-based approaches for polyphonic piano transcription using an architecture similar to [55], but making use of up-to-date training techniques, regularisers and taking into account hyper-parameter tuning. The 'ConvNet' architecture from the work of [33] can be seen in Fig. 24.4.

An influential work that used CNNs for multiple fundamental frequency estimation in polyphonic music was the *deep salience* representation proposed by Bittner et al. [12]. Contrary to most methods in AMT that produce a binary output, the model of [12] produces a non-binary time-pitch representation at 20 cent pitch resolution, which can be useful for both AMT applications but also for several downstream applications in the broader field of MIR. A particular contribution of this work was the use of a Harmonic Constant-Q Transform (HCQT) as input representation; the HCQT is a three-dimensional representation over frequency, time and the harmonic index, produced by computing several versions of the CQT by scaling the minimum frequency used by a harmonic. Figure 24.5 shows the pitch salience representation for the Mozart excerpt of Fig. 24.1, computed using the deep salience method of [12].

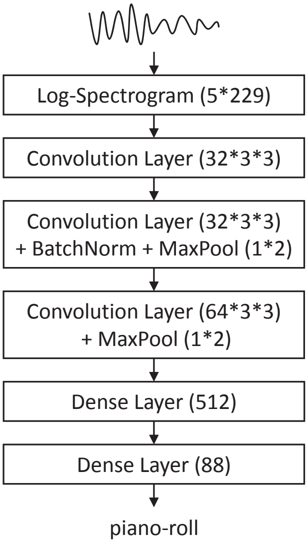
Log-Spectrogram (5*229)
Convolution Layer (32*3*3)
Convolution Layer (32*3*3)
+ BatchNorm + MaxPool (1*2)
Convolution Layer (64*3*3)
+ MaxPool (1*2)
Dense Layer (512)
Dense Layer (88)
piano-roll

Fig. 24.4 Model architecture for the convolutional neural network used in [33] for polyphonic piano transcription. The depicted network corresponds to the 'ConvNet' architecture of [33]

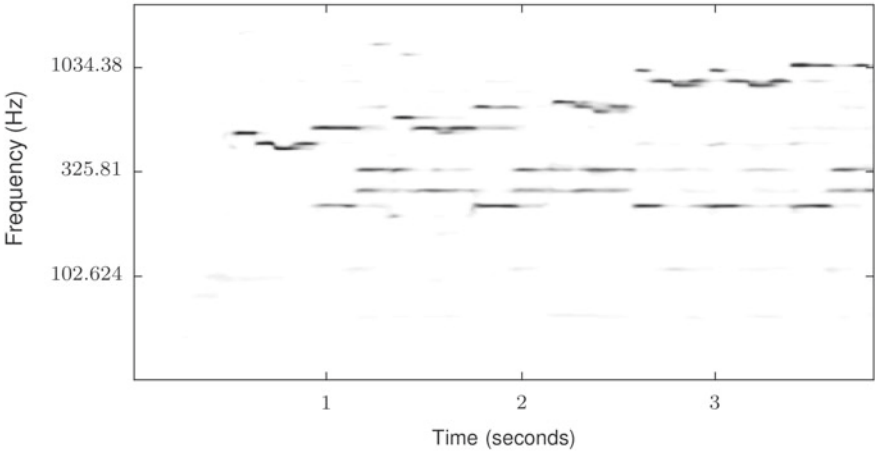

Fig. 24.5 Pitch salience representation for the excerpt of Fig. 24.1, using the deep salience method of [12]

The ability of CNNs in learning features in time or time-frequency representations keeps them still active in the AMT literature. This includes the work of Thickstun et al. [58] that was carried out as part of the MusicNet dataset, and compared feedforward and convolutional networks learned on raw audio inputs, as opposed to having a time-frequency representation as input. It is worth noting however that convolutional, and more broadly neural networks, when trained for AMT as a multi-label classification task, face the issue that they appear to learn combinations of notes exposed to them during training, and are not able to generalise unseen combinations of notes—the so-called *entanglement problem* as discussed in [34].

24.4.3 Multi-task Learning Methods

Recent research in machine learning has focused on *multi-task learning* [52], where multiple learning tasks are addressed jointly, thus exploiting task similarities and differences. In the context of AMT, multi-task learning has been shown to improve transcription performance in certain cases. Tasks related to AMT such as note level transcription, onset detection, melody estimation, bass line prediction and multi-pitch detection (also sharing similar chroma and rhythm features) can be integrated into one model that would exploit task interdependencies.

In the 'Onsets and Frames' system by Hawthorne et al. [28], which is currently considered the benchmark in automatic piano transcription, the authors used a deep Convolutional Recurrent Neural Network (CRNN) to jointly predict onsets and multiple pitches. The onset detection results are fed back into the model for further improving frame-wise multi-pitch predictions. The Onsets and Frames model was further improved in the work of Kim and Bello [35], which addresses the problem of expressing inter-label dependencies through an adversarial learning scheme.

Bittner et al. [11] proposed a multi-task model that jointly estimates outputs for several AMT-related tasks, including multiple fundamental frequency estimation, melody, vocal and bass line estimation. The authors show that the more tasks included in the model, the higher the performance and that the multi-task model outperforms the single-task equivalents. In another recent work [32], the authors designed a multi-task model with CNNs which enables four different transcription subtasks: multiple-f0 estimation, melody estimation, bass estimation and vocal estimation. Results on the method of [32] showed an overall improvement in the multi-task model compared to single task models.

24.4.4 Music Language Models

Inspired by work in the field of speech processing, where many systems for Automatic Speech Recognition (ASR) benefit from language models that predict the occurrence of a word or phoneme [31], researchers in MIR have recently attempted to use Music Language Models (MLMs) and combine them with acoustic models in order to improve AMT performance. While the problem of polyphonic music prediction using statistical machine learning models (such as n-grams and hidden Markov models) is not trivial, the emergence of neural network methods for high-dimensional sequence prediction has enabled the use of MLMs for polyphonic music.

One of the first works to use neural network-based MLMs for polyphonic music prediction and combine them with multi-pitch detection, was carried out by Boulanger-Lewandowski et al. [15]. The MLM was based on a combination of a recurrent neural network with a Neural Autogressive Distribution Estimator (NADE). The same RNN-NADE music language model was also used in [55], which was combined with a CNN as the acoustic model, showing that the inclusion of an MLM can improve transcription performance.

It was shown however that the MLMs which operate at the level of a small time frame (e.g. 10 msec) are only able to produce a smoothing effect in the resulting transcription [63]. More recently, Wang et al. [60] used an LSTM-RBM language model as part of their proposed transcription system, but each frame corresponds to an inter-onset interval as opposed to a fixed temporal duration, resulting in improved transcription performance when using note-based metrics. Finally, Ycart et al. [65] combined an LSTM-based music language model with a feedforward neural blending model which combines the MLM probabilities with the acoustic model probabilities. In line with past observations, the blending and language models work best when musically-relevant time steps are used (in this case, time steps corresponding to a 16th note).

24.4.5 Complete Transcription

Recent works have paid attention to *complete transcription*, where systems are developed to convert music audio into a music score. There are two common ways in designing a complete transcription system. A traditional way is by using a combination of several methods and subtasks of AMT to form an system that can transcribe music audio to a notation level, which usually involves estimating a piano-roll representation in an intermediate process [43]. Another way which has become increasingly popular is designing an end-to-end system that directly converts input audio or a time-frequency representation into a score level representation such as textual encoding, without having a piano-roll or similar intermediate representation in the pipeline. In this scenario, a deep learning network is used to link the system input and output. A challenge in designing a end-to-end system is that the input and output of the system cannot be aligned directly (one is a time-based representation and the other is a representation in terms of metres or symbolic encoding). As a result, research has focused on encoder-decoder architectures [16,46] which do not rely on framewise aligned annotations between the audio and music score.

A work worth mentioning which combined subtasks to build a transcription system is by Nakamura et al. [43]. In this work, the authors divided a whole transcription system into a stream of subtasks: multi-pitch analysis, note tracking, onset rhythm quantisation, note value recognition, hand separation and score typesetting. The final system reads a spectrogram calculated from music audio, and outputs readable music

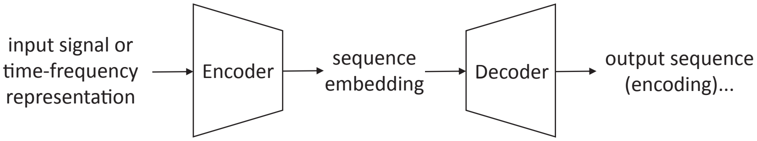

Fig. 24.6 General structure for an end-to-end AMT system using encoder-decoder architecture

scores. Offering the whole system structure, the authors did not focus on integrating algorithms for all the subtasks, but optimised methods for multi-pitch detection and rhythm quantisation. The improved subtask performance ends up adding to the final performance of the system.

Encoder-decoder mechanisms have also been used for AMT in recent years, with the advantage in creating complete transcription systems without estimating and integrating complicated subtasks. In Fig. 24.6, we provide an encoder-decoder structure commonly used in AMT systems. Recent works have showed the potential of encoder-decoder methods, although their performance on polyphonic music transcription remains less explored in the literature. In 2017, Carvalho and Smaragdis proposed a method for end-to-end music transcription using a sequence-to-sequence architecture combined with CNNs and RNNs [16]. The developed system can output a textual music encoding in Lilypond language from an input audio waveform. However, the work focused mainly on monophonic music (which showed high-level performance), but only a simple scenario of polyphonic music was tested (with two simultaneous melodies within a pitch range of two octaves). Another exploration on singing transcription by Nishikimi et al. [46] also used a sequence-to-sequence model. A point worth mentioning is that they applied an attention loss function for the decoder, which improved the performance of the singing transcription system. The work, still, focused only on monophonic singing voice.

Using an encoder-decoder architecture is a simple way of designing end-to-end AMT systems, but there are also other works using Connectionist Temporal Classification (CTC). A recent example is by Román et al. [51], in which the authors combined the use of a CRNN and a CTC loss function. The CTC loss function enables the system to be trained using pairs of the input spectrogram and output textual encoding. In that work, a simple polyphonic scenario is considered where four voices are included in a music piece (in string quarters or four-part Bach chorales). Still, the problem of end-to-end complete music transcription with unconstrained polyphony is still open.

24.5 Challenges

Although AMT is still very active as a topic within MIR, the performance of current AMT systems is still far from satisfactory, especially when it comes to polyphonic music, multiple instruments, non-Western music and ‘complete’ transcription. There are plenty of challenges in this area where further exploration is required. In this section, we summarise current challenges and provide potential further directions.

24.5.1 Datasets

The lack of annotated datasets is an aspect that limits the development of AMT systems. Due to the difficulty in collecting and annotating music recordings, there is still a lack of data for most music transcription tasks, especially for non-western music and certain musical instruments. Apart from the lack of large datasets, current datasets for AMT also have some limitations. For example, the temporal precision of annotations for some datasets with real recordings is not always satisfactory—which is also a reason that most AMT systems set a relatively large onset/offset *tolerance* for note tracking tasks. Also, dataset annotations are typically limited to note pitch, onset and offset times and sometimes note velocity. Additional annotations are needed for a more comprehensive transcription, such as rhythm, key information and expressiveness labels.

Recently, an increasing number of datasets has been released, which are based on synthesising MIDI files. MIDI files provide a good reference for multi-pitch detection since they provide temporally precise note annotations, but there are also limitations, since MIDI files do not provide annotations for score level transcription. Another limitation for synthesised data is that they might not reflect the recording and acoustic conditions of real-world audio recordings and can cause bias during model training.

24.5.2 Evaluation Metrics

Current evaluation metrics mainly focus on frame-wise and note-wise evaluations, where transcription results are provided in a piano-roll representation or note sequences. Benchmark evaluation metrics also do not model different error types beyond measuring precision and recall. For example, an extra note may be more severe than a missing note in a polyphonic music, on-key notes may be less noticed than off-key ones, and an error in a predominant voice may be more obvious compared to a similar error in a middle voice. Besides, much less work can be found in evaluating complete transcription systems.

There is also a lack of perceptual considerations in commonly used evaluation metrics. Some work [43,48] has attempted to create different types of errors, however these metrics still do not account for human perception. Deniel et al. provided an early work on perceptually-based multi-pitch detection evaluation [21], but is not widely used in the community. In addition, there is still no work on perceptually-based evaluation metrics for score-level transcription.

24.5.3 Non-Western Music

Most AMT methods aim specifically at modelling Western tonal music, but there is much less work done on automatically transcribing music cultures beyond Western tonal music, such as world, folk and traditional music. This results in AMT systems not being able to accurately or adequately transcribe non-Western music.

Differences between Western and non-Western music cultures that can affect the design of AMT systems include but are not limited to pitch space organisation and microtonality, the presence of heterophony (vs. homophony or polyphony occurring in Western tonal music), complex rhythmic and metrical structures, differences in tuning and temperament, differences in musical instruments and differences in methods for expressive performance and music notation amongst others. Despite the above differences, the lack of large annotated datasets is another limitation for music transcription research for non-Western music cultures.

24.5.4 Complete Transcription

Although research in AMT has increasingly been focusing on complete transcription in recent years, current methods and systems are still not suitable for general-purpose audio-to-score transcription of multi-instrument polyphonic music. Some systems for complete transcription rely on typesetting methods as a final step, but most typesetting methods assume a performance MIDI or similar representation as input and are not designed to take noisy input into account. In addition, when many tasks are combined into a whole system for complete transcription, the errors in each step can accumulate and worsen the system's performance. As for end-to-end transcription methods, current research is still limited to monophonic music and special cases for polyphonic music, mostly using synthetic audio. There is still a large room for further work towards the development of systems for complete music transcription.

24.5.5 Expressive Performance

Including expressive performance annotations is another challenge in current AMT research. Most AMT systems transcribe music into a defined framework of note pitch, onset and offset in a metre constrained format, but cover little expressive labels such as note velocity, speed symbols, as well as expressive playing techniques. It is currently hard to predict such information in AMT, although MIR research has been focusing on specific problems within the broader topic expressive music performance modelling (e.g. vibrato detection). How to incorporate the estimation and modelling of expressive performance into AMT systems remains an open problem.

24.5.6 Domain Adaptation

Due to the increasing use of synthesised datasets, or due to the mainstream use of piano-specific datasets for AMT, the ability of such models to generalise to real recordings, different instruments, acoustic recording conditions or music styles has become a problem worth considering. There is currently no research focusing on this question in the context of AMT, although the broader problem of *domain adaptation* has been attracting increasing interest in MIR and the broader area of machine learning.

For example, tasks in MIR such as music alignment and singing voice separation were explored in a recent paper [37] using domain adaptation methods based on variational autoencoders. We believe that similar domain adaptation methods can be applied to AMT tasks to solve existing problems such as the lack of data for some less popular instruments and dealing with the differences between synthesised and real-life recorded datasets or different recording conditions.

24.6 Conclusions

AMT is a core problem in the field of Music Information Retrieval (MIR), and has attracted a lot of attention during the past few decades. In this chapter, we review and discuss some of the main topics within the problem of AMT. We make a concrete definition of the problem of AMT, and describe the main subtasks in the AMT process (see Sect. 24.2). We also introduce the problem of complete transcription, which refers to the process of converting music audio into a music score representation. We review commonly used datasets and evaluation metrics for AMT (see Sect. 24.3), and look into the state of the art methodologies used in AMT (see Sect. 24.4). Current research on AMT has focused on methods using neural networks with promising results. We look into several topics in particular, including the use of commonly used neural network architectures, the use of multi-task learning methods, the use of music language models and methods for complete transcription. However, challenges still exist in the field of AMT, as we discussed in Sect. 24.5. A large room for improvement is open in areas such as building better datasets and evaluation metrics, building systems for non-Western music transcription, complete transcription, adding expressive performance in transcription results and considering domain adaptation. Given our review in this chapter, we believe that AMT is an open and promising field within both MIR and the broader intersection of music and artificial intelligence.

Acknowledgements L. Liu is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music and is supported by a China Scholarship Council and Queen Mary University of London joint Ph.D. scholarship. The work of E. Benetos was supported by RAEng Research Fellowship RF/128 and a Turing Fellowship.

References

1. Music Information Retrieval Evaluation eXchange (MIREX). Retrieved April 29, 2020, from http://music-ir.org/mirexwiki/.
2. SyncRWC dataset. Retrieved April 29, 2020, from https://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/SyncRWC/.
3. Bay, M., Ehmann, A. F., & Downie, J. S. (2009). Evaluation of multiple-F0 estimation and tracking systems. In *10th International Society for Music Information Retrieval Conference* (pp. 315–320). Kobe, Japan.
4. Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. (2005). A tutorial on onset detection of music signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *13*(5), 1035–1047.
5. Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2019). Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, *36*(1), 20–30. https://doi.org/10.1109/MSP.2018.2869928.
6. Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., & Klapuri, A. (2013). Automatic music transcription: Challenges and future directions. *Journal of Intelligent Information Systems*, *41*(3), 407–434. https://doi.org/10.1007/s10844-013-0258-3.
7. Benetos, E., & Holzapfel, A. (2013). Automatic transcription of Turkish makam music. In *14th International Society for Music Information Retrieval Conference* (pp. 355–360). Curitiba, Brazil.
8. Benetos, E., & Weyde, T. (2015). An efficient temporally-constrained probabilistic model for multiple-instrument music transcription. In *16th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 701–707). Malaga, Spain.
9. Bittner, R., & Bosch, J. J. (2019). Generalised metrics for single-f0 estimation evaluation. In *Proceedings of the 20th International Society of Music Information Retrieval Conference, ISMIR* (pp. 738–745). Delft, Netherlands.
10. Bittner, R., Salamon, J., Tierney, M., Mauch, M., Cannam, C., & Bello, J. (2014). Medleydb: A multitrack dataset for annotation-intensive mir research. In *International Society for Music Information Retrieval Conference* (pp. 155–160). Taibei, Taiwan.
11. Bittner, R. M., McFee, B., & Bello, J. P. (2018). *Multitask learning for fundamental frequency estimation in music*. arXiv:1809.00381 [cs.SD].
12. Bittner, R. M., McFee, B., Salamon, J., Li, P., & Bello, J. P. (2017). Deep salience representations for f0 estimation in polyphonic music. In *International Society for Music Information Retrieval Conference* (pp. 63–70). Suzhou, China.
13. Böck, S., & Schedl, M. (2012). Polyphonic piano note transcription with recurrent neural networks. In: *IEEE International Conference on Audio, Speech and Signal Processing* (pp. 121–124). Kyoto, Japan.
14. Bosch, J. J., Marxer, R., & Gómez, E. (2016). Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. *Journal of New Music Research*, *45*(2), 101–117.
15. Boulanger-Lewandowski, N., Bengio, Y., & Vincent, P. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *29th International Conference on Machine Learning*. Edinburgh, Scotland, UK.
16. Carvalho, R. G. C., & Smaragdis, P. (2017). Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 151–155).
17. Chen, C. H. (2016). *Handbook of Pattern Recognition and Computer Vision* (5th ed.). River Edge, NJ, USA: World Scientific Publishing Co., Inc.
18. Cogliati, A., & Duan, Z. (2017). A metric for music notation transcription accuracy. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 407–413).

19. Cogliati, A., Duan, Z., & Wohlberg, B. (2017). Piano transcription with convolutional sparse lateral inhibition. *IEEE Signal Processing Letters*, *24*(4), 392–396.
20. Cogliati, A., Temperley, D., & Duan, Z. (2016). Transcribing human piano performances into music notation. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 758–764).
21. Daniel, A., Emiya, V., & David, B. (2008). Perceptually-based evaluation of the errors usually made when automatically transcribing music. In *International Society for Music Information Retrieval Conference, ISMIR* (pp. 550–555).
22. Duan, Z., Pardo, B., & Zhang, C. (2010). Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 2121–2133.
23. Duan, Z., & Temperley, D. (2001). *Note-level music transcription by maximum likelihood sampling*.
24. Emiya, V., Badeau, R., & David, B. (2010). Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(6), 1643–1654.
25. Gòmez, E., & Bonada, J. (2013). Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing. *Computer Music Journal*, *37*(2), 73–90. https://doi.org/10.1162/COMJ_a_00180.
26. Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2003). RWC music database: Music genre database and musical instrument sound database. In *International Conference on Music Information Retrieval*. Baltimore, USA.
27. Hartmann, W. M. (1996). Pitch, periodicity, and auditory organization. *The Journal of the Acoustical Society of America*, *100*(6), 3491–3502. https://doi.org/10.1121/1.417248.
28. Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., et al. (2018). Onsets and frames: Dual-objective piano transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR* (pp. 50–57). Paris, France.
29. Hawthorne, C., Stasyuk, A., Robers, A., Simon, I., Huang, C.Z.A., Dieleman, S., et al. (2019). Enabling factorized piano music modeling and generation with the maestro dataset. In *International Conference on Learning Representations (ICLR)*.
30. Humphrey, E. J., Durand, S., & McFee, B. (2018). OpenMIC-2018: An open dataset for multiple instrument recognition. In *19th International Society for Music Information Retrieval Conference* (pp. 438–444). Paris, France.
31. Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing* (2nd ed.). Pearson.
32. Kelz, R., Böck, S., & Widmer, G. (2019). Multitask learning for polyphonic piano transcription, a case study. In *International Workshop on Multilayer Music Representation and Processing (MMRP)* (pp. 85–91). https://doi.org/10.1109/MMRP.2019.8665372.
33. Kelz, R., Dorfer, M., Korzeniowski, F., Böck, S., Arzt, A., & Widmer, G. (2016). On the potential of simple framewise approaches to piano transcription. In *Proceedings of International Society for Music Information Retrieval Conference* (pp. 475–481).
34. Kelz, R., & Widmer, G. (2017). An experimental analysis of the entanglement problem in neural-network-based music transcription systems. In *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*.
35. Kim, J. W., & Bello, J., Adversarial learning for improved onsets and frames music transcription. In *International Society for Music Information Retrieval Conference*.
36. Klapuri, A., & Davy, M. (Eds.). (2006). *Signal processing methods for music transcription*. New York: Springer.
37. Luo, Y. J., & Su, L. (2018). Learning domain-adaptive latent representations of music signals using variational autoencoders. In *Proceedings of International Society for Music Information Retrieval Conference* (pp. 653–660). Paris, France.
38. Manilow, E., Wichern, G., Seetharaman, P., & Roux, J. L. (2019). Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, NY.

39. McLeod, A., & Steedman, M. (2018). Evaluating automatic polyphonic music transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR* (pp. 42–49). Paris, France
40. McLeod, A., & Yoshii, K. (2019). Evaluating non-aligned musical score transcriptions with mv2h. In *Extended Abstract for Late-Breaking/Demo in International Society for Music Information Retrieval Conference, ISMIR*.
41. McVicar, M., Santos-Rodríguez, R., Ni, Y., & Bie, T. D. (2014). Automatic chord estimation from audio: A review of the state of the art. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(2), 556–575. https://doi.org/10.1109/TASLP.2013.2294580.
42. Molina, E., Barbancho, A. M., Tardòn, L. J., & Barbancho, I. (2014). Evaluation framework for automatic singing transcription. In *International Symposium on Music Information Retrieval Conference* (pp. 567–572).
43. Nakamura, E., Benetos, E., Yoshii, K., & Dixon, S., *Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization*.
44. Nakamura, E., Yoshii, K., & Sagayama, S. (2017). Rhythm transcription of polyphonic piano music based on merged-output hmm for multiple voices. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(4), 794–806.
45. Nam, J., Ngiam, J., Lee, H., & Slaney, M. (2011). A classification-based polyphonic piano transcription approach using learned feature representations. In *12th International Society for Music Information Retrieval Conference* (pp. 175–180). Miami, Florida, USA.
46. Nishikimi, R., Nakamura, E., Fukayama, S., Goto, M., & Yoshii, K. (2019). Automatic singing transcription based on encoder-decoder recurrent neural networks with a weakly-supervised attention mechanism. In *Proceedings of IEEE International Conference on Acoustics, Apeech and Signal Processing*.
47. Piszczalski, M., & Galler, B. A. (1977). Automatic music transcription. *Computer Music Journal*, *1*(4), 24–31.
48. Poliner, G., & Ellis, D. (2007). A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, *8*, 154–162.
49. Raffel, C. (2016). *Learning-based methods for comparing sequences, with applications to audio-to-midi alignment and matching*. Ph.D. thesis, Columbia University.
50. Rebelo, A., Fujinaga, I., Paszkiewicz, F., Marcal, A. R. S., Guedes, C., & Cardoso, J. S. (2012). Optical music recognition: State-of-the-art and open issues. *International Journal of Multimedia Information Retrieval*, *1*(3), 173–190. https://doi.org/10.1007/s13735-012-0004-6.
51. Román, M. A., Pertusa, A., & Calvo-Zaragoza, J. (2019). A holistic approach to polyphonic music transcription with neural networks. In *Proceedings of the 20th International Society for Music Information Retrieval Conference* (pp. 731–737). Delft, Netherlands.
52. Ruder, S. (2017). *An overview of multi-task learning in deep neural networks*. arXiv:1706.05098.
53. Salamon, J., Gomez, E., Ellis, D., & Richard, G. (2014). Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, *31*(2), 118–134. https://doi.org/10.1109/MSP.2013.2271648.
54. Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., et al. (2013). *Roadmap for music information research*. Creative Commons BY-NC-ND 3.0 license.
55. Sigtia, S., Benetos, E., & Dixon, S. (2016). An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(5), 927–939. https://doi.org/10.1109/TASLP.2016.2533858.
56. Smaragdis, P., & Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 177–180). New Paltz, USA.
57. Su, L., & Yang, Y. H. (2015). Escaping from the abyss of manual annotation: New methodology of building polyphonic datasets for automatic music transcription. In *International Symposium on Computer Music Multidisciplinary Research*.

58. Thickstun, J., Harchaoui, Z., & Kakade, S. M. (2017). Learning features of music from scratch. In *International Conference on Learning Representations (ICLR)*.
59. Virtanen, T., Plumbley, M. D., & Ellis, D. P. W. (Eds.). (2018). *Computational analysis of sound scenes and events*. Springer.
60. Wang, Q., Zhou, R., & Yan, Y. (2018). Polyphonic piano transcription with a note-based music language model. *Applied Sciences*, **8**(3). https://doi.org/10.3390/app8030470.
61. Wu, C., Dittmar, C., Southall, C., Vogl, R., Widmer, G., Hockman, J., et al. (2018). A review of automatic drum transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(9), 1457–1483. https://doi.org/10.1109/TASLP.2018.2830113.
62. Xi, Q., Bittner, R. M., Pauwels, J., Ye, X., & Bello, J. P. (2018). Guitarset: A dataset for guitar transcription. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR* (pp. 453–460). Paris, France.
63. Ycart, A., & Benetos, E. (2017). A study on LSTM networks for polyphonic music sequence modelling. In *18th International Society for Music Information Retrieval Conference (ISMIR)* (pp. 421–427).
64. Ycart, A., & Benetos, E. (2018). A-MAPS: Augmented MAPS dataset with rhythm and key annotations. In *19th International Society for Music Information Retrieval Conference Late Breaking and Demo Papers*. Paris, France.
65. Ycart, A., McLeod, A., Benetos, E., & Yoshii, K. (2019). Blending acoustic and language model predictions for automatic music transcription. In *20th International Society for Music Information Retrieval Conference (ISMIR)*.
66. Yu, D., & Deng, L. (Eds.). (2015). *Automatic Speech Recognition: A Deep Learning Approach*. London: Springer.

Lele Liu is a research student at the Centre for Doctoral Training in Artificial Intelligence and Music (AIM), based at the School of Electronic Engineering and Computer Science of Queen Mary University of London, UK. Her research is on automatic music transcription with neural networks. E-mail: lele.liu@qmul.ac.uk.

Emmanouil Benetos is Senior Lecturer at the School of Electronic Engineering and Computer Science of Queen Mary University of London and Turing Fellow at the Alan Turing Institute, in the UK. His research focuses on signal processing and machine learning methods for music and audio analysis, as well as applications to music information retrieval, sound scene analysis, and computational musicology. E-mail: emmanouil.benetos@qmul.ac.uk.