A Survey on Automatic Music Transcription

PRANAV BHAGWAT[1], VISHWAJIT SHELKE[2], AKHILESH MURUGKAR[3], KRISHIV DAKWALA[4], DR. SHWETA C. DHARMADHIKARI[5]

[1,2,3,4]Department of Information Technology, Pune Institute of Computer Technology

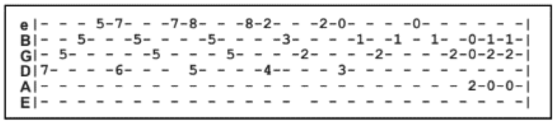
[5] Associate Professor, Pune Institute of Computer Technology

Abstract—Automatic Music Transcription (AMT) is a critical but less investigated problem in the field of music information retrieval. In this paper, we study different approaches for achieving Automatic Music Transcription using various methods based on pitch, timbre and note detection. Use of Convolutional Neural Network (CNN) and/or Long Short Term Memory Network (LSTM) is made to transcribe notes from the audio input. We also discuss source separation as a precursor to AMT and different approaches for the same.

Indexed Terms—Automatic Music Transcription (AMT), Pitch Detection, Note Detection, Deep Learning, Convolutional Neural Network (CNN), Long Short Term Memory Network (LSTM), Source Separation

I. INTRODUCTION

Automatic Music Transcription (AMT) is a highly lucrativefield with several approaches already existing. It deals withthe problem of converting an audio file containing music intoany musical notation format, be it score, sheet, tablatures, orothers. It can be said that AMT is analogous to the AutomaticSpeech Recognition (ASR), as both involve converting audiosignals into some symbolic notation [1].



```
e|- - - 5-7- - -7-8- - -8-2- - -2-0- - - -0- - - - - -|
B|- - 5- - -5- - - -5- - - -3- - - -1- -1 - 1- -0-1-1-|
G|- 5- - - - -5 - - - 5- - - -2- - - -2- - - -2-0-2-2-|
D|7- - - -6- - - 5- - - -4-- - - 3- - - - - - - - - -|
A|- - - - - - - - - - - - - - - - - - - - - - 2-0-0-|
E|- - - - - - - - - - - - - -  - - - - - - - - - - - -|
```

Fig. 1. ASCII Guitar Tablature of “Stairway to Heaven” [2]

AMT is a challenging task due to the variety of sub-tasks it contains such as pitch and note detection, harmonics detection, separation multiple instruments in the audio and the final task of actually transcribing the music. Various attempts have been made in this field from the use of genetic algorithm [2], latent harmonic allocation [3], and different architectures and combinations of Convolutional Neural Networks (CNN) and Long Short Term Memory Networks (LSTM).

Within Automatic Music Transcription itself, there are different types as follows:

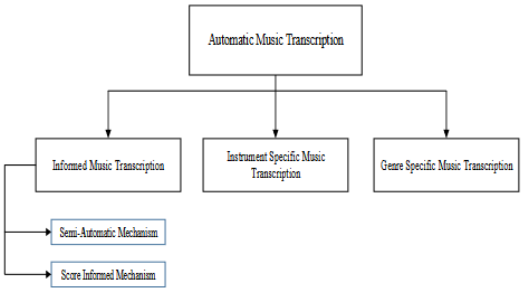

Automatic Music Transcription
Informed Music Transcription
Instrument Specific Music Transcription
Genre Specific Music Transcription
Semi-Automatic Mechanism
Score Informed Mechanism

Fig. 2. Classification of Automatic Music Transcription Methods [4]

This paper will focus on Informed Music Transcription and Instrument Specific Music Transcription, which is similar to a Semi-Automatic Mechanism as the knowledge of which instrument is creating the sound is already known i.e., it is informed.

However, it is not always necessary that the audio files contain a single instrument only, they can contain an ensemble or noise or a vocal track along with an instrument; this is known as polyphonic music. When it comes to polyphonic music, source separation is a key step to be able to differentiate between the different instruments within the audio [5]. Here, the timbre, i.e., quality of each unique instrument comes into play. For transcription purposes, accurate pitch and note detection is key due to the harmonics created by the instrument and in some cases even the environment it is being played in. The fundamental pitch helps establish the notes accurately which can then be represented into the chosen notation.

This paper reviews the work done in the a fore mentioned fields, mainly Automatic Music Transcription (AMT) but also source separation, supervised learning using score and other techniques used as modifications or necessary pre-processing steps.

II. LITERATURE SURVEY

A. On Source Separation

Y. N. Hung et al. [6] proposed the use of musical score as a supervised learning technique to identify and separate the sounds of different musical instruments. They used a transcriptor to provide both temporal and frequency information to a separator that used the score as a weak label for the target.

A. Jansson et al. [7] proposed the application of U-Net
Convolutional Neural Network architecture previously used only in medical imaging for source separation of audio, namely the separation of the vocal and backing track in music. The model is trained on a triplet of original track, instrumental track and vocal track using a custom dataset. The model was primarily compared against the Chimera model which uses deep clustering on the iKala and MedleyDB datasets where it provided state of the art performance.

L. Lin et al. [8] proposed a zero shot unified model for
source separation, transcription and synthesis of new pieces. They used an encoder-decoder combination in the form of U-Net Convolutional Neural Network to implement Pitch-Timbre disentanglement which allows for the implementation of all 3 tasks.

E. Manilow, P. Seetharaman, and B. Pardo [9] proposed a single deep learning architecture that can both separate an audio recording of a musical mixture into constituent single-instrument recordings and transcribe these instruments into a human-readable format at the same time.

Y.-N. Hung and A. Lerch [10] proposed a novel multitask structure to investigate using instrument activation information to improve source separation performance. Furthermore, they investigate their system on six independent instruments, a more realistic scenario than the three instruments included in the widely used MUSDB dataset, by leveraging a
combination of the Medley DB and Mixing Secrets datasets. The results show that their proposed multitask model out performs the baseline Open-Unmix model on the mixture of Mixing Secrets and MedleyDB dataset while maintaining comparable performance on the MUSDB dataset.

K. A. Pati and A. Lerch [11] explores the problem of automatically detecting electric guitar solos in rock music. A baseline study using standard spectral and temporal audio features in conjunction with an SVM classifier is carried out. To improve detection rates, custom features based on predominant pitch and structural segmentation of songs are designed and investigated. The evaluation of different feature combinations suggests that the combination of all features followed by a post-processing step results in the best accuracy. A macro-accuracy of 78.6% with a solo detection precision of 63.3% is observed for the best feature combination.

D. Règnier, N. Martin, and L. Bigo [12] proposed a
computational method to identify rhythm guitar sections in symbolic tablatures. They defined rhythm guitar as sections that aim at making the listener perceive the chord progression that characterizes the harmony part of the song. A set of31 high level features is proposed to predict if a bar in atablature should be labeled as rhythm guitar or not.

B. On Informed Music Transcription

Wu et al. [13] proposed a U-net style neural network architecture with the purpose of identifying different musical instruments in an audio clip. Here the bottleneck layer that connects the encoder and decoder contains self-attention blocks using Image transformer to better capture the temporal modeling of the audio data. The output of the model is a multi-channel representation where each channel is a time-frequency image of the extracted individual instruments present.

R. Tuohy and W. Potter [2] proposed the use of a Genetic Algorithm to generate staff notation tablature for guitar music. The generator accepts as an input some representation of the note sequence that defines a piece of music and generates a tablature as output. This approach is useful as it considers the large sample space of different positions present on the guitar fret board and selects the ones that are easiest to
play in succession.

K. Tanaka et al. [14] proposed a 3 step approach for Multi-Instrument Multi-Pitch Estimation (MI-MPE). This includes extracting pitchgram and spectrogram from an audio signal using Constant Q transform and Short Term Fourier Transform, a 3 layer Bidirectional LSTM to generate masks for the timbre space and piano roll which can be applied to the pitchgram and spectrogram to achieve an estimated piano roll i.e. the pitch of the instrument over time. Deep Spherical Clustering using K-Means is used to cluster the similar instruments together.

F. Simonetta, S. Ntalampiras, and F. Avanzini [15] proposed a method which benefits from HMM-based score-to-score alignment and AMT, showing a remarkable advancement beyond the state-of-the-art. They designed a systematic procedure to take advantage of large data sets which do not offer an aligned score. Finally, they performed athorough comparison and extensive tests on multiple datasets.

N. Takahashi, N. Goswami, and Y. Mitsufuji [16] proposed a novel architecture that integrates long short-term memory (LSTM) in multiple scales with skip connections to efficiently model long-term structures within an audio context. The experimental results show that the proposed method outperforms MMDense Net, LSTM and a blend of the two networks. The number of parameters and processing time of the proposed model are significantly less than those for simple blending. Furthermore, the proposed method yields better results than those obtained using ideal binary masks for a singing voice separation task.

I. Barbancho, L. J. Tardon, S. Sammartino, and A. M. Barbancho [17] proposed a system for the extraction ofthe tablature of guitar musical pieces using only the audio wave form. The analysis of the in harmoni city relations between the fundamentals and the partials of the notes played is the main process that allows to estimate both the notes played and the string/fret combination that was used to produce that sound. A procedure to analyze chords will also be described. This procedure will also make use of the in harmoni city analysis to find the simultaneous string/fret combinations used to play each chord. The proposed method is suitable for any guitar type: classical, acoustic, and electric guitars. The system performance has been evaluated on a series of guitar samples from the RWC instruments database and our own recordings.

E. Mistler [18] proposed two different architectures. In the first approach, guitar frettings are directly predicted based onpreviously played frettings. This is accomplished by a LongShort-Term Memory Recurrent Neural Network, enhanced bya notion of intention of the musical outcome. The second approachpredicts the difficulty of a fretting in terms of a cost function,rather than predicting the ideal fretting directly. The costfunction is based on conditional probabilities in the trainingdata and estimated by a Feed-forward Neural Network.

Table I. A Summary of Literature Survey

Sr. No.	Paper Title	Major Deliverable	Shortcomings	Dataset	Metrics Used
1.	Transcription is all you need: Learning to separatemusicalmixtureswithscore as supervision [6]	Use of score for transcriptor supervision learning (piano	Only considermixtures ofacoustic piano,distortedelectri	Slakh2100 [19]	Scale-invariant-signal-to-distortion

		roll)	c guitar,and electric bass		ratio (SI-SDR)
2.	A unified model for zero-shot music for source separation, transcription and synthesis [8]	Pitch-Timbre Disentanglement module using encoder-decoder	Only considers classical music ensemble	URMP [5]	Frequency vs Time Spectrograms
3.	Audio-to-score alignment using deep automatic music transcription [15]	Use of stacked and parallel CNN + GRU (Gated Recurrent Units) combinations	Unreliable for non-piano solo music	Custom dataset created by mixing multiple datasets	Matching offset and onset predictions
4.	Identification of rhythm guitar sections in symbolic tablatures [12]	Use of LSTM classifier to get F1 Score of 0.95	The definition of foreground and background instruments is not standardized	Custom annotated dataset using GuitarPro	Recall, Precision, F1 Score
5.	Multi-Instrument Music Transcription Based on Deep Spherical Clustering of Spectrograms and Pitchgrams [14]	Able to deal with undefined musical instruments with low error	Misestimation in creation of piano roll	Slakh2100 [19]	Precision, Recall, F-measure
6.	Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments [9]	Achieves automatic music transcription on polyphonic music	Can separate only a few instruments out of all present	Slakh [19], MAPS [9], GuitarSet [20]	Recall, Precision, F1 Score
7.	Multitask learning for instrument activation aware music source separation [10]	Equal or better separation quality than the baseline Open-Unmix model	Fewer training samples cause poor performance with too many instruments present	MUSDB-HQ and a combination of Mixing Secrets with MedleyDB	Source to distortion ratio (SDR), source to interference ratio (SIR), source to artifact ratio (SAR), and Image to Spatial distortion Ratio (ISR)
8.	Multi-instrument automatic music transcription with self-attention-based	Outperforms the baseline methods on	Occurrence of overfitting	MAESTRO and MusicNet	Recall, Precision, F1 Score

	instance segmentation [13]	frame-level MPS (multi-pitch streaming)			
9.	Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation [16]	MMDenseLSTM outperforms a naïve combination of BLSTM and MMDenseNet despite having much fewer parameters.	Performance for bass is low	DSD100 and MUSDB18	Source to distortion ratio (SDR)
10.	A dataset and method for guitar solo detection in rock music [11]	Good accuracy using SVM classifier on standard spectrogram	Dataset is very small and results cannot be generalized	Custom made GSD (Guitar Solo Detection) dataset	Micro-Accuracy, Macro-Accuracy, Precision, Recall, F-measure and Specificity
11.	Singing voice separation with deep u-net convolutional networks [7]	Novel application of the U-Net architecture for source separation	Focused on singing voice, separation from other voices not mentioned	Custom dataset	Source-to-Distortion Ratio (SDR), Source-to-Interference Ratio (SIR), and Source-to-Artifact Ratio (SAR)
12.	Inharmonicity-based method for the automatic generation of guitar tablature [17]	The detection of the note-string-fret combinations is done without any previous knowledge of the played notes	Increasing error rate as the number of notes simultaneously played in the chords increases	Custom Dataset	Manual comparison of note/chord detected
13.	Generating guitar tablatures with neural networks [18]	The approach used does not focus on the finger position of the guitarist but on the best fitting frets and strings which avoids implicit customization of hand sizes	The dataset used is small and needs more data points to improve the model	Converted GuitarPro files to MusicXML	Distance Heuristic

14.	A genetic algorithm for the automatic generation of playable guitar tablature [2]	A note can be played on many different positions on the fretboard of a guitar, and these can be explored using a genetic algorithm	Neither the achieved accuracy nor heuristic used is explicitly mentioned	GuitarPro files	Unknown

III. FUTURE PROSPECTS

Most of the work and datasets (whether they be audio files or multi-modal i.e., video files) created for the same, whether they be on a solo instrument or polyphonic, they have emphasized on using classical music which uses piano, violin and other instruments but notably ignores one of the most popular instruments known around the world: the guitar. Approaches using the guitar have not simultaneously achieved thorough understanding of the tablatures as well as playability i.e., the ease of playing the tablature on the guitar by considering criteria such as fret spread and hand shape. Furthermore, the work done in AMT primarily uses western music theory a sits base, similar to how advancements in Natural Language Processing were primarily done in English before achieving the same results in languages such as Hindi, French, German and others. Thus, there is scope for such work to be done in Indian music which uses “Sa-Re-Ga-Ma-Pa-Dha-Ni” as its base to notate music.

CONCLUSION

In this paper, we considered the problem of Automatic Music Transcription and have reviewed the methods so far proposed by several authors for transcribing the music with different information available for the transcription and classified those works under Informed Music Transcription and Instrument Specific Music Transcription. We also reviewed various source separation approaches which have been used as an essential pre-processing step to extract a particular instrument’s audio to then generate the transcription.

Many papers suggested multiple approaches utilizing deep learning such as CNNs and/or LSTMs, while others implemented genetic algorithm or other machine learning algorithms. Their deliverables and shortcomings have been highlighted to show a comprehensive view of the work done so far. However, the problem of Automatic Music Transcription has not yet been solved and the field shows promise for further exploration.

REFERENCES

[1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, “Automatic music transcription,” IEEE Signal Processing Magazine, vol. 1053, no. 5888/19,2019.

[2] D. R. Tuohy and W. D. Potter, “A genetic algorithm for the automaticgeneration of playable guitar tablature,” in ICMC, pp. 499–502, 2005.

[3] K. Yazawa, D. Sakaue, K. Nagira, K. Itoyama, and H. G. Okuno,“Audio-based guitar tablature transcription using multipitch analysisand playability constraints,” in 2013 IEEE International Conference onAcoustics, Speech and Signal Processing, pp. 196–200, IEEE, 2013.

[4] B. Gowrishankar and N. U. Bhajantri, “An exhaustive review of automatic music transcription techniques: Survey of music transcription techniques,” in 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), pp. 140–152,IEEE, 2016.

[5] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, “Creating amultitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications,” IEEE Transactions on Multimedia, vol. 21, no. 2, pp. 522–535, 2018.

[6] Y.-N. Hung, G. Wichern, and J. Le Roux, “Transcription is all you need: Learning to separate musical mixtures with score as supervision,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 46–50, IEEE, 2021.

[7] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, “Singing voice separation with deep u-net convolutional networks,” 2017.

[8] L. Lin, Q. Kong, J. Jiang, and G. Xia, “A unified model for zero-shot music source separation, transcription and synthesis,” arXiv preprintarXiv:2108.03456, 2021.

[9] E. Manilow, P. Seetharaman, and B. Pardo, “Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 771–775, IEEE,2020.

[10] Y.-N. Hung and A. Lerch, “Multitask learning for instrument activation aware music source separation,” arXiv preprint arXiv:2008.00616, 2020.

[11] K. A. Pati and A. Lerch, “A dataset and method for guitar solo detection in rock music,” in Audio Engineering Society Conference: 2017AES International Conference on Semantic Audio, Audio Engineering Society, 2017.

[12] D. Règnier, N. Martin, and L. Bigo, “Identification of rhythm guitar sections in symbolic tablatures,” in International Society for Music Information Retrieval Conference (ISMIR 2021), 2021.

[13] Y.-T. Wu, B. Chen, and L. Su, “Multi-instrument automatic music transcription with self-attention-based instance segmentation,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28,pp. 2796–2809, 2020.

[14] K. Tanaka, T. Nakatsuka, R. Nishikimi, K. Yoshii, and S. Morishima, “Multi-instrument music transcription based on deep spherical clustering of spectrograms and pitchgrams.,” in ISMIR, pp. 327–334, 2020.

[15] F. Simonetta, S. Ntalampiras, and F. Avanzini, “Audio-to-score alignment using deep automatic music transcription,” in 2021 IEEE 23rdInternational Workshop on Multimedia Signal Processing (MMSP), pp. 1–6, IEEE, 2021.

[16] N. Takahashi, N. Goswami, and Y. Mitsufuji, “Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation,” in 2018 16th International workshop on acoustic signal enhancement (IWAENC), pp. 106–110, IEEE, 2018.

[17] I. Barbancho, L. J. Tardon, S. Sammartino, and A. M. Barbancho, “In harmonicity-based method for the automatic generation of guitar tablature,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 6, pp. 1857–1868, 2012.

[18] E. Mistler, “Generating guitar tablatures with neural networks,” Master of Science Dissertation, The University of Edinburgh, 2017.

[19] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, “Cutting music source separation some slakh: A dataset to study the impact of training data quality and quantity,” in 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),pp. 45–49, IEEE, 2019.

[20] Q. Xi, R. M. Bittner, J. Pauwels, X. Ye, and J. P. Bello, “Guitarset: Adataset for guitar transcription.,” in ISMIR, pp. 453–460, 2018.