Vol.7 No.1 Jan 2024

Multimedia Research

Received 12 November, Revised 24 November, Accepted 15 December

A Comprehensive Review on Automatic Music Transcription: Survey of Transcription Techniques

Sagar Latake

Citiustech Healthcare Private Limited, Bengaluru, Karnataka, India
splatake@gmail.com

Abstract: "In music, transcription is the practice of notating a piece or a sound which was previously unnotated and/or unpopular as a written music". An absolute transcription will be performed only when the timing, pitching, and instruments of all sound events are solved. In music transcription systems, a MIDI file is found to be a suitable format for melodic notations. This survey intends to make a review of 65 papers that concern music transcription using machine learning techniques. Accordingly, systematic analyses of the adopted techniques are carried out and presented briefly. The performances and related maximum achievements of each contribution are also portrayed in this survey. Moreover, the various datasets used in music transcription techniques were considered and reviewed in this work. Finally, the survey portrays the research problems and weaknesses that may be supportive for researchers to introduce the latest techniques related to music transcription.

Keywords: Music Transcription; Machine Learning; Carnatic Music; Performance; HMM Model.

Nomenclature

Abbreviation	Description
AMT	Automatic Music Transcription
CQT	Constant Q Transform
CNN	Convolutional Neural Network
CRFs	Conditional Random Fields
CSC	Convolutional Sparse Coding
DNN	Deep Neural Networks
e2e	End-to-End
EM	Expectation-Maximization
FAR	False Alarm Rate
FPR	False Positive Rate
GA	Genetic Algorithm
GM	Gaussian Mixtures
HMMs	Hidden Markov Models
HALCA	Harmonic Adaptive Latent Component Analysis
HHSMM	Hierarchical Hidden Semi-Markov Model
HMRF	Hidden Markov Random Fields
MAE	Mean Absolute Error
MOR	Mean Overlap Ratio
MUSEMBLE	MUSicen EMBLE
NN	Neural Network
NMF	Non-Negative Matrix Factorisation
NMD	Non-Negative Matrix Decomposition
PLCA	Probabilistic Latent Component Analysis
RPA	Raw Pitch Accuracy
RAE	Relative Absolute Error
RCA	Raw Chroma Accuracy
RF	Random Forest
SNR	Signal-to-Noise Ratio
SAGE	Space-Alternating Generalized Expectation-Maximization
SVM	Support Vector Machines
TSPs	Time-Constrained Sequential Patterns
TPR	True Positive Rate
VCR	Voicing Recall Rate

https://doi.org/10.46253/j.mr.v7i1.a1

1. Introduction

"Music transcription is defined to be the act of listening to a piece of music and of writing down musical notation for the notes that constitute the piece, i.e. transforming an acoustic signal into a symbolic representation, which comprises notes, their pitches, timings, and a classification of the instruments used. AMT [1] [2] is the process of automatically converting a musical sound signal into its representation as musical notation, through digital analysis of the musical signal". The AMT was the objective of many researchers from the time of its establishment, and currently, it has covered a wider range of subtasks [3] [4]. The appliances of AMT consist of automated recovery of melodic data, interactive music systems, and musicological examination. Moreover, it was viewed as one of the most important technologies that facilitates theories in musical signal processing [5] [6].

In previous years, machine learning was introduced that tries to imitate Human learning processes. This field of analysis focuses on adding up cognition skills to machines by modeling the learning process [7]. Machine Learning and AMT [11] [12] are associated since the intention of both is to make computers perceive and understand music on their own. So far, numerous techniques have been introduced for addressing AMT tasks based on supervised learning approaches, in which different frameworks such as NN, CNN, SVM, and so on were exploited that ensure better results.

Accordingly, the musical tradition of India was separated into two huge traditions, such as, "Hindustani and Carnatic music" [8] [9] [10]. Hindustani classic music was mainly practiced in the northern part of India and it has been originated from prehistoric Persian, Vedic, and various folk customs. Whereas Carnatic music is pure Indian music and it is mostly practiced in the southern part of India. Thus, the music of South India ("Sanskrit, Karnataka Sangitam) is referred to as Carnatic or Karnātak music in English". The main contribution of this paper is given as follows.

1. Presents an extended review on music transcription using machine learning techniques and analyzes the varied techniques adopted in each paper.
2. Makes a review of performance measures and their corresponding maximum achievements in each work.
3. Carries out an analysis of various datasets adopted for music transcription using machine learning techniques.

This paper deals with the following sections: Section 2 illustrates the related works based on music transcription techniques. The wide-ranging review of adopted techniques, performances, and maximum attainments is preferred in section 3. The evaluation of adopted datasets for music transcription using machine learning techniques is represented in Section 4. Furthermore, Section 5 organized the research gaps and challenges, and the conclusion is portrayed in Section 6.

2. Literature Review

2.1 Related Works

2.1.1 Polyphonic Piano Music

In 2018, Yukawa and Hideaki [1] suggested a convex-analytic model for music transcription to NMF using the Moreau envelope. In addition, the optimization issue was numerically resolved by the proximal backward–forward splitting technique that does not require any secondary variables. At last, the outcome of the method had shown better effectiveness and minimal errors when compared to other traditional methods. In 2013, Orti *et al.* [3] implemented a realistic signal decomposition technique with single-pitch and harmonicas parameters using instrument-based information. Accordingly, the basic operations were learned with a supervised procedure that obtained spectral prototypes for diverse melodic instruments. At last, the experimental results have shown the superiority of the proposed method in terms of accuracy and runtime.

In 2008, Rho *et al.* [5] developed a new music retrieval system called MUSEMBLE based on 2 different characteristics: (i) A hummed or sung query was transcribed automatically with enhanced accuracy (ii) A user inquiry was remodeled using GA that improved the retrieval process. At last, the proposed method revealed better transcription accuracy than other traditional methods. In 2016, Cogliati *et al.* [8] demonstrated a new technique for the transcription of piano music automatically in a context-based setting. This technique employed CSC for estimating the music signals as the summary of dictionary elements. At last, the comparative analysis of the method provided better time precision and accuracy than other existing models.

In 2017, Rizzi *et al.* [17] executed an AMT approach that depends on a supervised NMD model. In the initial stage, the NMD was carried out and in the subsequent stage; threshold filtering was deployed for

refining the notes. Finally, the experimental outcomes indicated the competitiveness of the adopted method against the conventional methods regarding accuracy. In 2006, Bello *et al.* [28] developed a technique that grouped spectral data in the frequency domain and exploited a rule-oriented approach for dealing with the well-known issues of harmonicas and polyphony. From the analysis, precise time-domain transcription was achieved from the simulated outcomes.

In 2011, Cho and CC [35] implemented a source-oriented dictionary model for proficient music representation, which involved three steps. Initially, the fundamental components of musical waveforms were decomposed. Subsequently, the Gabor atoms were prioritized and exploited for synthesizing novel atoms to make up a compacted dictionary. In 2011, Durrieu *et al.* [36] devised a source/filter signal approach, which was successfully exploited within a major instrument separation system and melody extraction system. Both these approaches attained better results at numerous international assessment operations. In 2010, Klapuri and Tuomas [39] developed an effective algorithm for modeling and indicating time-varying melodic sounds. Here, individual sounds were observed and the specific class to which the song belongs will be determined. Eventually, the experimental results have shown better computational complexity and accurate outcomes than other existing models.

In 2011, Grindlay and Daniel [43] applied a probabilistic model for transcribing music recordings that contain several polyphonic instrument sources. The approach modeled the individual tools and thus, the detected notes were assigned to their relevant sources. At last, the proposed approach revealed better performance in terms of F-measure. In 2018, Akbari *et al.* [53] examined a real-time learning-oriented model for visual transcription of piano music using SVM - CNN classification. The entire procedure in this method depends on the visual examination of the piano keyboard and fingers and hands of the pianist. The method had proved to have better accuracy and F1 score than other traditional methods. In 2014, Serrano *et al.* [55] analyzed the "monophonic constrained signal decomposition model" applied to polyphonic waveforms that involved numerous monophonic resources from diverse melodic instruments. The choral parameter was predominantly effectual for tonal devices; since every note was related to a unique source.

In 2015, Arora and Behera [57] developed a model, where a short-time magnitude spectrum was decomposed using PLCA for F0 estimation. Moreover, HMRF was exploited for clustering F0s into diverse sources. In the end, better recall and precision were attained by the adopted model when compared to existing models. In 2018, Jose *et al.* [60] introduced a machine learning based approach, which aimed at inferring note tracking automatically for piano music. Here, every pitch band was segmented into unique instances that were classified as non-active or active note events. Finally, the proposed approach provided a better F-measure than other traditional models. In 2016, Tiago *et al.* [62] introduced an unsupervised method, which allowed the discovery of a threshold that depends on well-known characteristics of the auditory signal. Furthermore, source codes and instructions were provided, which allowed the reproduction of all portrayed experimentations. In 2014, Arı *et al.* [63] developed a proficient approach based on exemplar selection and randomized matrix decomposition, which easily handled huge dictionary matrixes in real appliances. The adopted method was further applied for transcribing polyphonic piano music.

2.2.2 Jazz

In 2020, Miguel *et al.* [2] developed an e2e technique depending on DNN for audio-to-score transcription of music from monophonic quotes. Here, an audio file was given as input that was modeled as a frame sequence, and a DNN was trained to provide a sequence of encoded music notes. Finally, the proposed method has shown better outcomes than the other traditional methods in terms of error.

2.2.3 World

In 2015, Akbari and Cheng [9] established the "computer vision-oriented automated music transcription model (claVision)" for performing piano musical transcription. Rather than processing the musical audio, the system performed transcription only from the video captivated by a camera. Finally, the approach provided better outcomes than other traditional models in accuracy and latency. In 2016, O'Hanlon *et al.* [10] explored the subspace modeling of note spectrum that was employed for coupling activation of associated atoms into pitched subspace. Moreover, the NMF model was exploited for tuning the choral subspace dictionary, which led to enhanced NMF-oriented AMT outcomes.

In 2010, Bertin *et al.* [15] introduced experimental and theoretical outcomes regarding constrained NMF in a Bayesian model. Subsequently, a SAGE approach was exhibited for estimating the parameters. In 2010, Rao and Preeti [46] introduced a model, where the temporal volatility of tone harmonics was deployed for identifying the pitch of voice. Results demonstrated that the modelled approach had recovered the lost musical information and it also outperformed other existing tune extraction systems. In 2012, Marolt [48] presented a model, which estimated the count of bells in a soundtrack and their estimated spectrum. The model exploited a customized edition of the k-means algorithm. For

transcribing a recording, a probabilistic framework was proposed, which integrated onset detection and factorization with former knowledge of bell sound rules.

In 2020, Shen *et al.* [51] presented a visual analysis approach that utilized auto-encoders to facilitate the investigation of conventional Chinese music. Moreover, the labeled dataset was constructed that was converted into spectrograms. Finally, the results revealed better outcomes using the adopted model over other traditional models. In 2008, Krige *et al.* [64] presented an automated transcription system of a particular song into note sequences. Accordingly, HMMs were deployed to represent both unique notes and the transitions among them for capturing the inconsistency of the approximated pitch. At last, the comparative analysis of this method has shown better results than other existing models in terms of accuracy. In 2009, Lavner and Dima [50] established a proficient approach for segmenting auditory signals into music or speech. The approach included learning and classification phases. Accordingly, an automated process was deployed for feature selection and the Bayesian model was exploited for carrying out classification.

2.2.4 Classical

In 2009, Costantini *et al.* [4] demonstrated a method for analyzing the music from a polyphonic piano, prompted by actions related to the notes. In this work, the classification of notes and offset detections was carried out based on CQT and SVMs. In 2016, Sigtia*et al.* [6] made an analysis based on the NN model for polyphonic piano musical transcription. Here, the auditory model was an NN that estimated the pitch probability in an audio frame. Finally, the experimental outcomes of the adopted method exposed better performance and reduced run-time. In 2011, Argenti *et al.* [14] implemented a technique that aimed at evaluating the intensity, duration, onset times, and pitch of synchronized sounds played by varied instruments. The adopted technique has revealed superior performance in multiple F0 tracking than other compared models. In 2017, Abeßer and Schuller [16] dealt with the automated transcription of guitar recordings. Further, SVM was deployed as the classifier for classifying the instrument-level constraints. In the end, the efficient implementation of the transcription system was demonstrated in terms of accuracy.

In 2017, Nakamura *et al.* [18] used an inference method that was suitable for any combined-output HMM. In addition, the impact of structural design and constraints of the technique were examined in terms of accuracy for voice separation and transcription of rhythm. In 2007, Poliner *et al.* [20] analyzed "how to define the tune of music audio, and what use it might be". This technique achieved around 70% accurate transcription at the frame level and it distinguished the absence or presence of the melody in the music. In 2013, Fuentes *et al.* [22] introduced the HALCA model that considered spectral and pitches of variations simultaneously. All the constraints were approximated via the EM approach. In 2017, Nakamura *et al.* [24] presented a numerical technique for musical transcription, which estimated offsets and onsets from score times of note. In addition, a context-tree model was constructed and it was combined with Markov random model. This method has the capability of automatically capturing of structure of voices effectively.

In 2008, Dubnov [33] presented a technique for analyzing the music structure, which captured global repetition and local prediction properties of auditory signals. In addition, spectral anticipation as well as recurrence analysis were carried out to find the temporal features of music. In 2012, Akira *et al.* [37] determined a technique for recuperating fingerings for a specified portion of violin music for recreating the tone of a known audio recording of the portion. This was accomplished by examining an auditory signal that determined the most probable sequence of 2D fingerboard location. In 2014, Arora and Laxmidhar [40] suggested a new method that deployed unsupervised and semi-supervised approaches for source clustering. The adopted hypothetical model was executed using graph clustering and PLC analysis. Finally, the results of the system indicated better accuracy than other existing models. In 2010, Paul *et al.* [42] introduced probabilistic generative models for the mutual modeling of spectral features such as excitations temporal activations, and harmonicity. Further, a general EM model was derived for collecting model constraints. In 2008, Gillet and Gaël [44] presented progressions on source separation and music transcription with a focal point on drum waveforms. Along with the proficient fusion policies; the transcribing system integrated a huge set of optimally elected features. In the end, better accuracy was accomplished by the adopted model over other schemes. In 2018, Dhara *et al.* [52] extracted music contour from a melody file using a salience-oriented melody extraction technique. Primarily, the notes were portrayed by optimizing the duration of music and tolerance band. Moreover, the outcome of this transcription system included notes, their period, and their offset and onset boundaries. In 2016, Cazau *et al.* [59] developed an automated transcription scheme devoted to the "marovany musical repertoires". The adopted transcription model included the PLCA algorithm, which was post-processed with HMM. At last, the proposed method revealed better outcomes than other existing models in terms of F-measure.

2.2.5 Western Music

In 2009, David [12] developed a model that combined three features ("metrical analysis, harmonic analysis, and stream segregation") of symbolic music analysis into one process, by which the composite interactions among the structures could be captured. Further, this technique also provided an approximation of the probability of note patterns and it also aided in the modelling of AMT. In 2008, Lee and Slaney [19] described an audio chord transcription model, which exploited representative data for training HMM and offered the best frame-level detection results. Moreover, musical keys were identified by selecting a key model via maximum likelihood that provided enhanced accuracy.

In 2012, Barbancho *et al.* [25] implemented a technique for deriving the finger configurations from a recorded guitar performance in an automatic manner. The technique was modeled as an HMM model and the acoustic features were attained from the frequency estimator. Moreover, the considered count of chords was considerably larger in this approach. In 2012, Reis *et al.* [29] presented a technique for multiple frequency (F0) evaluation on piano recording. Consequently, GA was exploited for analyzing the overlapping tones and also for searching the certain F0 combinations. At last, the quality of the results of the method was proved in terms of computing time and correlation. In 2013, Stanislaw *et al.* [38] developed a group of probabilistic representative polyphonic pitch frameworks, which accounted for both the "vertical" and "horizontal" pitch structures. Moreover, the capability of the frameworks to forecast pitch data was computed in terms of the "contextual cross-entropy" model. At last, the experimental analysis has shown efficient effective outcomes to accuracy.

In 2012, Rao *et al.* [47] investigated the usage of signal sparsity for analyzing the lengths of the window. In addition, diverse metrics of sparsity deployed to the local band, like the Gini index and kurtosis were computed. At last, the relative analysis of the designed method provided improved robustness when compared to other existing models. In 2011, Shen and Lee [56] implemented an interactive "Whistle-to-Music composing system" that enabled users to compose MIDI music by whistling into a microphone. Moreover, the user could make their practice using computer-aided compositions like chord generation and melodic deviation.

2.2.6 Classic and Jazz

In 2011, Benetos and Dixon [13] suggested a technique for automatic transcription of musical waveforms depending on combined multiple-F0 evaluation. For selecting an optimal pitch combination, a score function was exploited and accordingly, HMMs and CRFs were deployed for classification purposes. In the end, the proposed approach provided better outcomes than other traditional models in terms of accuracy and minimal error. In 2012, Benetos and Dixon [23] developed a probabilistic model for multiple-instrument based AMT. Moreover, the shift-invariant feature of the technique was deployed for recognizing the changes in frequency modulations, tunes, and pitch content visualizes. Finally, the experimental analyses have exposed minimal error outcomes for the adopted method.

In 2004, Marolt [26] presented a connectionist scheme for automated transcription of polyphonic piano music. In addition, groups of partials were tracked by the combination of adaptive oscillators into NN. At last, the outcomes using the system have represented a feasible substitute for existing transcription approaches. In 2015, Carrillo and Marcelo [41] suggested a new indirect acquirement technique for estimating constant violin controls with a database of formerly recorded violin performance. The numerical techniques deployed were HMM with Multivariate GM. At last, the analysis of the approach indicated superior performance with high accuracy at a lower cost.

In 2010, Anglade *et al.* [45] presented a genre classification model using both high-level and low-level harmony features. In addition, the timbre features were extended via RF, which classified genre classes. At last, the comparative analysis of the method provided better classification rates than other existing models. In 2012, Ren and Jyh [49] implemented the usage of TSPs as effectual features for classifying music genres. Initially, an automated language detection method was carried out for tokenizing every music piece into the HMM index. At last, SVM was deployed for carrying out accurate classification tasks. In 2010, Cañadas *et al.* [58] demonstrated a multiple-F0 evaluation model for automated polyphonic transcription of music. The adopted model searched for fundamental frequencies, which reduced the spectral distance at every auditory frame. Moreover, HMM was exploited which offered better accuracy and improvement. In 2015, Sun &Hongyan [61] described a technique for transcribing the major polyphonic music structure automatically by examining music tonality associated with musicological information. In addition, an approach was used for transcribing the most important factor of polyphonic tune from the viewpoint of tonality.

2.2.7 Pop

In 2016, Kroher and Gómez [7] extracted the melody and applied a new contour filtering procedure for eliminating the pitch contour segments that originated from the guitar. The adopted scheme outperformed conventional music transcription models concerning onset detection and accuracy. In 2013,

Gómez and Bonada [21] dealt with automated transcription of flamenco musical recordings, particularly, a cappella lyrics. Initially, the information of flamenco lyrics was studied and a transcription system was proposed based on energy estimation and frequency. At last, the method indicated better outcomes than other existing models.

2.2.8 Rock Music

In 2020, Nishikimi *et al.* [30] adopted a Bayesian HHSMM, which integrated a melodic score approach for examining the rhythms and local keys of musical notes. Eventually, this method revealed better outcomes when compared with other existing models. In 2009, Paulus and Klapuri [65] implemented a technique for drum transcription from polyphonic music through HMMs. The intention was to identify the sequential locations of un-pitched percussive tunes and thus, the instrumentals played were recognized. Finally, the proposed technique using HMMs has revealed better outcomes than other existing models.

2.2.9 Indian Music

In 2020, Bhalarao and Raval [54] suggested an automatic tabla syllable transcription technique utilizing image processing. Consequently, supervised classification was exploited for labeling every stroke depending on its image for a certain syllable. At last, the comparative analysis of the method provided a better F1 score than other existing models.

2.2.10 Others

In 2015, Wan *et al.* [11] developed a new piano-based transcription approach that deployed both visual and audio features. The contribution included 2 parts: An onset recognition technique and a computer-vision technique were presented for enhancing audio-only piano music transcription. This was carried out by tracking the positions of the pianist's hands on the piano. In 2016, Ewert and Sandler [27] developed a new waveform model, where temporal dependencies among spectral patterns were modeled that appeared similar to the features of HMM and NMF models. Finally, the experimental outcomes have revealed that the adopted method offered a higher F -measure than the compared schemes.

In 2017, Cogliati *et al.* [31] employed CSC together with lateral inhibition parameters for estimating a music signal. While transcription, the dictionary elements were preset and their sequential activations were approximated and post-processed for obtaining the note length, onset, and pitch estimation. Finally, the transcription accuracy was found to be improved by this method. In 2008, Wang and Bingjun [32] developed musical-instrument training, a computer-assisted, scenario, where the client should practice with no human training for most of the time. The targeted users of the primary system began with singing students and violin students who were comfortable with computers. In 2010, Carabias *et al.* [34] suggested an unsupervised procedure for obtaining music scene-adaptive spectral patterns for every MIDI note. In addition, the attained harmonic dictionary was deployed to note-event recognition with harmonizing pursuit. At last, the experimental analysis of the approach revealed superior performance with a lower error rate and higher accuracy.

In 2022, Gao *et al.* [66] spotlighted transformer network architecture called the multi-transcriber through combined individual transcription to OMR and AMT model. Initially, a jointly supervised learning system was used to train an encoder-decoder pathway for each chord and lyric. Also, before decoding lyrics and chord signals independently, it adds a single encoder in the front end to benefit from correlations between them. Finally, the performance of all existing systems was tested against datasets to evaluate the accuracy of the results obtained. This approach was compared with other methods that were previously used. This model outperformed existing systems in terms of accuracy and speed.

In 2022, Wang, *et al.* [67] practiced the AST framework called Music YOLO for note-level transcription. Pitch tagging was first realized using a unique spectrogram peak search procedure. This required finding the spectrogram matrix's peak point to determine an exact pitch (f pitch). Based on note object detection's onset, offset, and bottom frequencies, this was done. Finally, Algorithm 2 outlined the specific steps that need to be taken during this process. The result was tested on various datasets and got good results in flamenco singing transcription and the melody contour feature.

In 2022, Wang *et al.* [68] utilized both the power subtraction method, where the noise was intended to be a component of the signal and maintained a Gaussian distribution with a standard deviation of vib, and a rule-based approach to combine nearby plucks from non-repeated strings into a strum interval. High-quality annotations were made possible by the initial optical sensor recording of a 4-track string vibration signal. After that, the TEAS was put into use to verify the dataset's scalability. The article incorporated and assessed three expressive analysis methods in this system. Lastly, the number of existing and fresh MIR duties was determined, and two AMT models were only examined for potential future research.

In 2022, Simonetta *et al.* [69] introduced a sensorized acoustical piano and AMT systems to resyn the size MIDI data. Initially, different contexts were simulated to evaluate the changes in interpretation

when the context was altered. The data was then meaningfully correlated with listener responses, bridging any gaps, and objectively measured using existing technologies such as AMT or MIDIs, among others. Additionally, multimodal machine learning techniques such as score-informed AMTs, wavelet transforms, paraconsistent feature engineering, and audio-to-score alignment methods were explored to develop new scores for music transcription purposes. The paper concluded that both MIDI and AMT were to be used to bridge the gap.

In 2022, Wu *et al.* [70] presented a semi-supervised method using low-rank matrix factorization systems to achieve this task. For training, it first needed tagged recordings of each note, which was accomplished using low-rank matrix factorization methods like nonnegative matrix factorization. As a result, optimization problems were solved using multiplicative algorithms that alternate between W and h updates while preserving non negativity and ensuring cost reduction with the help of zero padding when necessary. Finally, adaptive peak-picking methods were used along with the CNMF activation method on the MAPS dataset for note F4 from a song to learn single-note templates after 500 outer iterations had been performed.

In 2022, Holzapfel *et al.* [71] contrasted two areas, namely the examination of these transcriptions by experts and computers. Initially, the transcription of eighteen transcribers was collected and evaluated with computational transcription through music transcription (AMT) methods. Metrics were included, such as comparing lists of notes in terms of physical time such as pitch onset and offset or binary "piano-roll" representations. The output result showed a high correlation between metrics and human ratings.

3. Wide-Ranging Review of Adopted Techniques, Performances, and The Maximum Attainments

3.1 Review of Adopted Techniques

The review of adopted techniques in each work is discussed in this section, and the illustration in Fig. 1. From the review, it was noted that the NMF method was used in [1] [3] [10] [34] [43] [48] [55] and [62]. HMM system was adopted in [7] [13] [18] [19] [20] [25] [27] [36] [41] [49] [58] [64] and [65]. In addition, SVM was adopted in [4] [16] [44] [52] [53] [60] and Otsu thresholding method was used in [9]. Markov random is used in [24]. Furthermore, optimization based methods like GA were exploited in [5] [29], and greedy algorithm was exploited in [35] respectively. Moreover, other clustering techniques such as the spectral clustering method were adopted in [33], K-means clustering was deployed in [4] [39], graph clustering was exploited in [40] and exemplar-based clustering was adopted in [63] correspondingly. In addition, the CSC was deployed in [8], [17] [31]. Consequently, the matched filtering was adopted in [11], and the LSTM Autoencoder method was deployed in [51]. Similarly, the NN [26] based models like CRNN were adopted in [2], RNN was deployed in [6] and ANN method was deployed in [61]. Also, the Bayesian approach was deployed in [12] [15] [30] [32] and [38]. The PLCA model was utilized in [23] [57] [59] and random forest was deployed in [45]. Viterbi algorithm was deployed by [46] and the Fourier transform was used in [14] [28] [47] and [56]. EM model was adopted in [22] [37] [42]. SAC algorithm, Decision-Tree, and Hough's transformation were adopted in [21] [50] and [54] respectively.

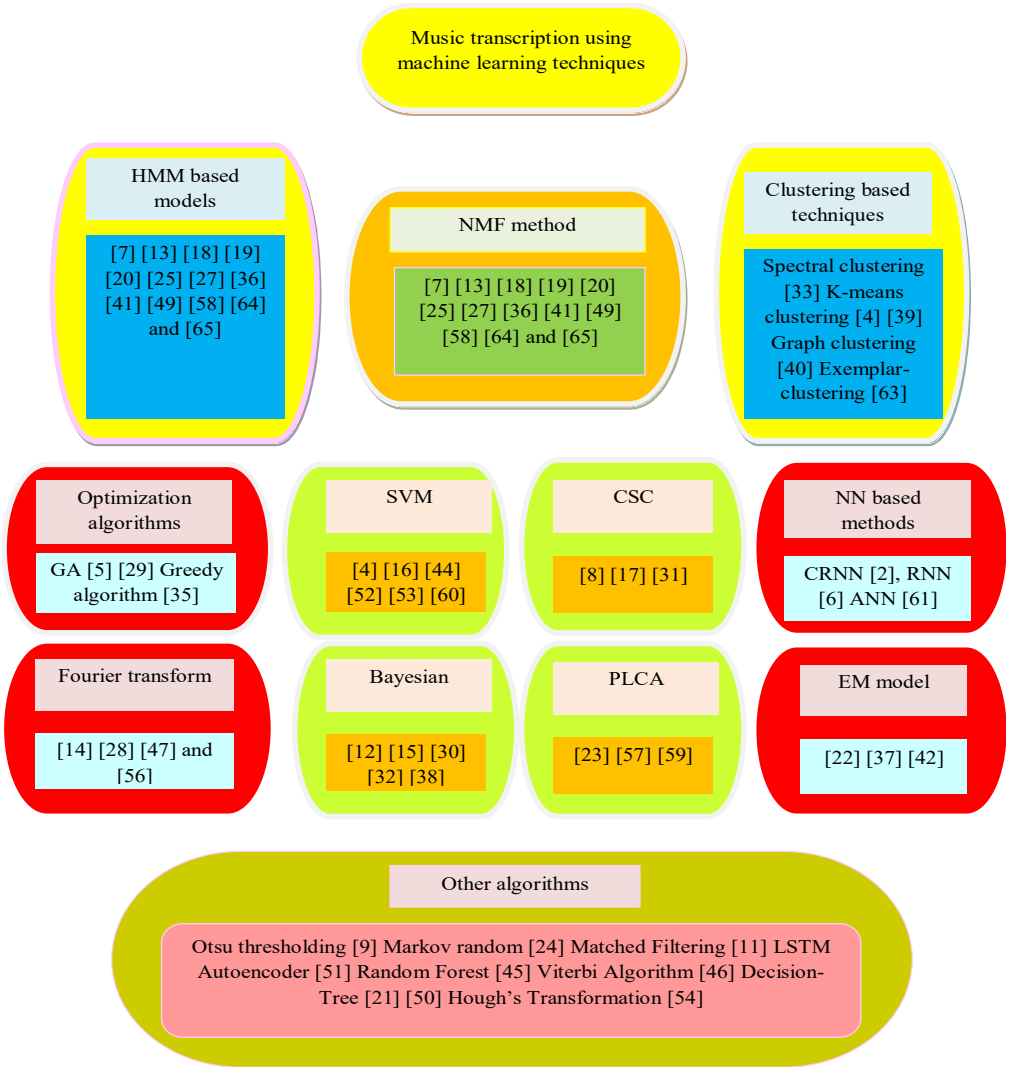

Fig.1. *Pictorial representation of music transcription using machine learning techniques*

3.2 Analysis of Performance Measures

Table I describes the performance measures that are adopted in various contributions of music transcription. From Table I, it is observed that 35 papers have made a performance analysis under accuracy that has contributed about 52.31% of the reviewed works, and the precision was analyzed in 29 papers that had contributed about 44.62% of the entire works. Likewise, the recall and F1-score have contributed about 43.08% and 6.15% (28 and 4 papers). Further, F-Measure and error values have been adopted in 47.69% (31 papers) and 20% (13 papers). Moreover, computational time, standard deviation, confidence interval, FAR, and Information gain have contributed about 3.08% (2 papers) of the entire contribution. On the other hand, the latency, MOR, and frequency have been adopted in 4.61% (3 papers). Accordingly, measures like RCA, TPR, FPR, decay score, SNR, MAE, RAE, frame length, pitch accuracies, chroma accuracy, percentage of total time correctly labeled and loss have contributed about 1.54% respectively. Similarly, the measures including the window length, VCR, and RPA have contributed 1.54%.

Table1. *Review on various performance measures in music transcription Models*

Measures	Citations
Accuracy	[3] [4] [5] [6] [7] [8] [9] [13] [14] [16] [17] [18] [19] [20] [21] [23] [24] [25] [26] [30] [32] [37] [40] [42] [45] [48] [49] [52] [54] [55] [58] [59] [60] [63] [64]
Precision	[4] [5] [6] [7] [8] [9] [10] [11] [13] [14] [15] [16] [17] [22] [23] [29] [27] [30] [31] [36] [38] [40] [44] [47] [48] [57] [59] [61] [63]
Recall	[4] [5] [6] [7] [8] [9] [10] [11] [13] [14] [15] [16] [17] [22] [23] [29] [27] [30] [31] [36] [38] [44] [47] [48] [57] [59] [61] [63]
F1-Score	[9] [53] [54] [60]
F-Measure	[4] [6] [7] [8] [10] [11] [13] [14] [15] [16] [17] [21] [22] [23] [24] [27] [29] [30] [31] [36] [38] [43] [44] [48] [52] [57] [59] [61] [62] [63] [65]
Error value	[1] [2] [5] [17] [18] [21] [22] [26] [29] [30] [47] [50] [58]
Computational time	[20] [49] [50] [55]
standard dev	[49] [58]
confidence interval	[3] [13]
Latency	[9] [50] [53]
MOR	[15] [31] [43]
Frequency	[47] [51] [56]
FAR	[16] [42]
Information gain	[49] [59]
	Miscellaneous measures
RCA	[16]
TPR	[28]
FPR	[28]
Decay Score	[29]
SNR	[39]
MAE	[41]
RAE	[41]
Frame length	[46]
Pitch Accuracies	[46]
Chroma accuracy	[46]
% of total time correctly labelled	[12]
Loss	[51]
window length	[1]
VCR	[16]
RPA	[16]

3.3 Analysis of Maximum Performance

The maximum performance obtained in each paper is represented in Table II. From the review, accuracy measured in [21] has obtained a better range of 100%, and precision in [27] has a higher value of 100%. Moreover, recall has obtained a better value of 100% and measured in [27] and the F1-score has obtained a better value of 0.98 and examined in [54] respectively. Similarly, error value, computational time, standard deviation, and confidence interval have attained better values of 1.2, 2.59sec, 11.5%,and 95% and it has been examined in [1] [49] [58] and [13] correspondingly. The measures such as latency, MOR, frequency, FAR, and information gain have attained higher values of 7.0 ms, 50, 20 kHz, 0.427, and 0.45 and they have been analyzed in [9] [15] [51] [16] and [49] respectively. Also, RCA, TPR, FPR, RPA, SNR, MAE, and RAE were exploited in [16] [28] [28] [16] [39] [41] and [41] and they have acquired higher values of 0.796, 80.9%, 14.7%, 0.765, 28Db, 0.09 and 23.2% correspondingly. In addition, the frame length, pitch accuracies, chroma accuracy, % of the total time correctly labeled, loss, window length, and VCR have attained higher values of 40ms, 92%, 90.2%, 80.8%, 0.001, 23.22 ms, and 0.890 and they have been measured in [46] [46] [46] [12] [51] [1] and [16] respectively.

Table 2. *Maximum performance attained in the Contributed works*

Sl. no	Citation	Performance measures	Maximum performance
1	[21]	Accuracy	100%
2	[27]	Precision	100%
3	[27]	Recall	100%
4	[54]	F1-Score	0.98
5	[27]]	F-Measure	100%
6	[1]	Error value	1.2
7	[49]	Computational time	2.59sec
8	[58]	standard dev	11.5%
9	[13]	confidence interval	95%
10	[9]	Latency	7.0 ms
11	[15]	MOR	50
12	[51]	Frequency	20 kHz
13	[16]	FAR	0.427
14	[49]	Information gain	0.45
15	[16]	RCA	0.796
16	[28]	TPR	80.9%
17	[28]	FPR	14.7%
18	[16]	RPA	0.765
19	[39]	SNR	28dB
20	[41]	MAE	0.09
21	[41]	RAE	23.2%
22	[46]	Frame length	40ms
23	[46]	Pitch Accuracies	92%
24	[46]	Chroma accuracy	90.2%
25	[12]	% of total time correctly labelled	80.8%
26	[51]	Loss	0.001
27	[1]	window length	23.22 ms
28	[16]	VCR	0.890

4. Analysis of Adopted Datasets for Music Transcription Using Machine Learning Techniques

4.1 Review of Music Transcription Datasets

The pictorial representation of the different datasets adopted for music transcription is given in Fig. 2. From the review, the RWC database was used in 14 works that offer about 22% of the total contribution. In addition, the MIDI database has been adopted by 18% of the entire works, and the MAPS database was used in 17% of the reviewed works. Likewise, ENST drums and MIR-1k were utilized in 3% of the reviewed papers. GTZAN and ISMIR04 contributed about 5% and other datasets like Western pop datasets were used in 12% of the entire works.Moreover, the PEDB database, IDMT-SMT-BASS-VAR-DUR, Oracle, woodwind, Poliner and Ellis, C2M and TON, Scheirer-Slaney, BACH, MIREX, and IMIDI have offered about 2% of the total contribution.

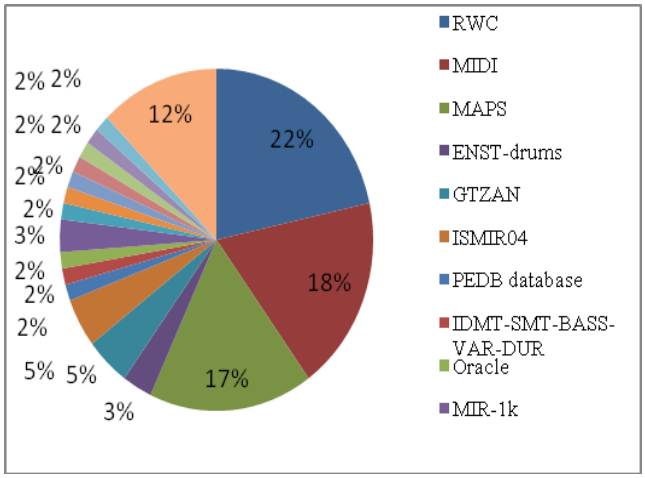

Fig.2. Pie chart representation on different datasets used for music transcription

5. Research Gaps and Challenges

"Musical transcription defines the process of converting a piece of music into some form of musical notation, which will display the musical notes played across time". For past decades, the issues of AMT have gained significant research attention owing to the several appliances related to varied areas like automatic search, musicological analysis, and so on. In melodic notations, few patterns include piano rolls, scores, or periodic sequences of musical tones [19] [20]. Even people with better musical training find it difficult to transcribe the notes manually after listening to a bit of music. One might encounter numerous problems when carrying out this task, like recognizing which instrument plays every note or identifying the beat of every note, however, the major challenge is to identify the pitch of notes.

Despite noteworthy developments in AMT analysis [12] [18], there are no end-user applications, which could reliably and accurately transcribe tune-containing genres and instrument combinations in music. The AMT issue is separated into numerous subtasks that take account of: "multi-pitch detection, note onset/offset detection, loudness estimation and quantization, instrument recognition, extraction of rhythmic information, and time quantization". The core difficulty of AMT is the evaluation of simultaneous pitches in a time frame, also termed multi-pitch or multiple-F0 detection. In certain conventional auditory transcription systems [20] [25], methods such as visual modality are deployed to assist music transcription. In a polyphonic mixture containing numerous instruments, the interference of concurrently occurring sounds is expected to limit the detection performance. Considering these drawbacks of AMT, further research is required to deploy it effectively.

6. Conclusion

This paper offered a complete review of music transcription using machine learning techniques. Here, various methodologies adopted in the reviewed works were analyzed and described. Furthermore, this survey analyzed the performance measures adopted in each paper.

- This paper is a review of AMT by analyzing 65 papers from different years. Initially, the techniques used are analyzed.
- The analysis has reviewed the performance measures and their maximum achievements from various machine learning techniques.
- Further, various datasets exploited in each reviewed work were also examined and determined using a pie chart.
- Finally, the paper presented a variety of research problems that may be helpful for researchers to carry out further work on music transcription.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] Masahiro Yukawa, Hideaki Kagami, "Supervised nonnegative matrix factorization via minimization of regularized Moreau-envelope of divergence function with application to music transcription", Journal of the Franklin Institute, vol. 355, no. 4, pp. 2041-2066, March 2018.

[2] Miguel A. Román, Antonio Pertusa, Jorge Calvo-Zaragoza, "Data representations for audio-to-score monophonic music transcription", Expert Systems with Applications, 30 December 2020.

[3] J. J. Carabias - Orti, F. J. Rodriguez-Serrano, P. Vera-Candeas, F. J. Cañadas-Quesada, N. Ruiz-Reyes, "Constrained non-negative sparse coding using learnt instrument templates for realtime music transcription", Engineering Applications of Artificial Intelligence, August 2013.

[4] Giovanni Costantini, Renzo Perfetti, Massimiliano Todisco, "Event based transcription system for polyphonic piano music" Signal Processing September 2009.

[5] Seungmin Rho, Byeong - jun Han, Eenjun Hwang, Minkoo Kim, "MUSEMBLE: A novel music retrieval system with automatic voice query transcription and reformulation", Journal of Systems and Software, July 2008.

[6] 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. doi: 10.1109/TASLP.2016.2533858.

[7] N. Kroher and E. Gómez, "Automatic Transcription of Flamenco Singing From Polyphonic Music Recordings," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 5, pp. 901-913, May 2016. doi: 10.1109/TASLP.2016.2531284.
[8] A. Cogliati, Z. Duan, and B. Wohlberg, "Context-Dependent Piano Music Transcription With Convolutional Sparse Coding," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 12, pp. 2218-2230, Dec. 2016. doi: 10.1109/TASLP.2016.2598305.
[9] M. Akbari and H. Cheng, "Real-Time Piano Music Transcription Based on Computer Vision," IEEE Transactions on Multimedia, vol. 17, no. 12, pp. 2113-2121, Dec. 2015. doi: 10.1109/TMM.2015.2473702.
[10] K. O'Hanlon, H. Nagano, N. Keriven, and M. D. Plumbley, "Non-Negative Group Sparsity with Subspace Note Modelling for Polyphonic Transcription," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 3, pp. 530-542, March 2016. doi: 10.1109/TASLP.2016.2515514.
[11] 11. Y. Wan, X. Wang, R. Zhou, and Y. Yan, "Automatic Piano Music Transcription Using Audio-Visual Features," Chinese Journal of Electronics, vol. 24, no. 3, pp. 596-603, 07 2015. doi: 10.1049/cje.2015.07.027.
[12] S. Tsuruta, M. Fujimoto, M. Mizuno, and Y. Takashima, "Personal computer-music system-song transcription and its application," IEEE Transactions on Consumer Electronics, vol. 34, no. 3, pp. 819-823, Aug. 1988. doi: 10.1109/30.20189.
[13] E. Benetos and S. Dixon, "Joint Multi-Pitch Detection Using Harmonic Envelope Estimation for Polyphonic Music Transcription", IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 6, pp. 1111-1123, Oct. 2011. doi: 10.1109/JSTSP.2011.2162394.
[14] F. Argenti, P. Nesi, and G. Pantaleo, "Automatic Transcription of Polyphonic Music Based on the Constant-Q Bispectral Analysis," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 6, pp. 1610-1630, Aug. 2011. doi: 10.1109/TASL.2010.2093894.
[15] N. Bertin, R. Badeau, and E. Vincent, "Enforcing Harmonicity and Smoothness in Bayesian Non-Negative Matrix Factorization Applied to Polyphonic Music Transcription," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, pp. 538-549, March 2010. doi: 10.1109/TASL.2010.2041381.
[16] J. Abeßer and G. Schuller, "Instrument-Centered Music Transcription of Solo Bass Guitar Recordings," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 9, pp. 1741-1750, Sept. 2017. doi: 10.1109/TASLP.2017.2702384.
[17] A. Rizzi, M. Antonelli, and M. Luzi, "Instrument Learning and Sparse NMD for Automatic Polyphonic Music Transcription," IEEE Transactions on Multimedia, vol. 19, no. 7, pp. 1405-1415, July 2017. doi: 10.1109/TMM.2017.2674603.
[18] E. Nakamura, K. Yoshii, and S. Sagayama, "Rhythm Transcription of Polyphonic Piano Music Based on Merged-Output HMM for Multiple Voices," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 4, pp. 794-806, April 2017. doi: 10.1109/TASLP.2017.2662479.
[19] K. Lee and M. Slaney, "Acoustic Chord Transcription and Key Extraction From Audio Using Key-Dependent HMMs Trained on Synthesized Audio," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 2, pp. 291-301, Feb. 2008. doi: 10.1109/TASL.2007.914399.
[20] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gomez, S. Streich, and B. Ong, "Melody Transcription From Music Audio: Approaches and Evaluation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1247-1256, May 2007. doi: 10.1109/TASL.2006.889797.
[21] E. Gómez and J. Bonada, "Towards Computer-Assisted Flamenco Transcription: An Experimental Comparison of Automatic Transcription Algorithms as Applied to A Cappella Singing," Computer Music Journal, vol. 37, no. 2, pp. 73-90, June 2013. doi: 10.1162/COMJ_a_00180.
[22] 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. doi: 10.1109/TASL.2013.2260741.
[23] E. Benetos and S. Dixon, "A Shift-Invariant Latent Variable Model for Automatic Music Transcription," Computer Music Journal, vol. 36, no. 4, pp. 81-94, Dec. 2012. doi: 10.1162/COMJ_a_00146.
[24] E. Nakamura, K. Yoshii, and S. Dixon, "Note Value Recognition for Piano Transcription Using Markov Random Fields," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 9, pp. 1846-1858, Sept. 2017. doi: 10.1109/TASLP.2017.2722103.
[25] A. M. Barbancho, A. Klapuri, L. J. Tardon, and I. Barbancho, "Automatic Transcription of Guitar Chords and Fingering From Audio," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 3, pp. 915-921, March 2012. doi: 10.1109/TASL.2011.2174227.
[26] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," IEEE Transactions on Multimedia, vol. 6, no. 3, pp. 439-449, June 2004. doi: 10.1109/TMM.2004.827507.
[27] S. Ewert and M. 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. doi: 10.1109/TASLP.2016.2593801.
[28] J. P. Bello, L. Daudet, and M. B. Sandler, "Automatic Piano Transcription Using Frequency and Time-Domain Information," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 6, pp. 2242-2251, Nov. 2006. doi: 10.1109/TASL.2006.872609.
[29] G. Reis, F. Fernandez de Vega, and A. Ferreira, "Automatic Transcription of Polyphonic Piano Music Using Genetic Algorithms, Adaptive Spectral Envelope Modeling, and Dynamic Noise Level Estimation," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2313-2328, Oct. 2012. doi: 10.1109/TASL.2012.2201475.

[30] R. Nishikimi, E. Nakamura, M. Goto, K. Itoyama, and K. Yoshii, "Bayesian Singing Transcription Based on a Hierarchical Generative Model of Keys, Musical Notes, and F0 Trajectories," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1678-1691, 2020. doi: 10.1109/TASLP.2020.2996095.
[31] Cogliati, Z. Duan and B. Wohlberg, "Piano Transcription With Convolutional Sparse Lateral Inhibition," IEEE Signal Processing Letters, vol. 24, no. 4, pp. 392-396, April 2017. doi: 10.1109/LSP.2017.2666183.
[32] Ye Wang and Bingjun Zhang, "Application-specific music transcription for tutoring", IEEE MultiMedia, Vol. 15, No. 3, pp. 70-74, 2008.
[33] Shlomo Dubnov, "Unified view of prediction and repetition structure in audio signals with application to interest point detection", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 2, pp. 327-337, 2008.
[34] J. J. Carabias-Orti, P. Vera-Candeas, F. J. Cañadas-Quesada, and N. Ruiz-Reyes, "Music scene-adaptive harmonic dictionary for unsupervised note-event detection", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 3, pp. 473-486, 2010.
[35] Namgook Cho and CC Jay Kuo, "Sparse music representation with source-specific dictionaries and its application to signal separation", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 2, pp. 326-337, 2011.
[36] Jean-Louis Durrieu, Barak David, and Guilhem Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation", IEEE Journal of Selected Topics in Signal Processing, Vol. 5, No. 6, pp. 1180-1191, 2011.
[37] Akira Maezawa, Katsutoshi Itoyama, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Automated violin fingering transcription through analysis of an audio recording", Computer Music Journal, Vol. 36, No. 3, pp. 57-72, 2012.
[38] Stanislaw Raczynski, Emmanuel Vincent, and Shigeki Sagayama, "Dynamic Bayesian networks for symbolic polyphonic pitch modeling", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21, No. 9, pp. 1830-1840, 2013.
[39] Anssi Klapuri and Tuomas Virtanen, "Representing musical sounds with an interpolating state model", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 3, pp. 613-624, 2010.
[40] Vipul Arora, and Laxmidhar Behera, "Musical source clustering and identification in polyphonic audio", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, No. 6, pp. 1003-1012, 2014.
[41] Alfonso Perez-Carrillo and Marcelo M. Wanderley, "Indirect Acquisition of Violin Instrumental Controls from Audio Signal with Hidden Markov Models", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 5, pp. 932-940, 2015.
[42] Paul H. Peeling, and Simon 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, 2010.
[43] Graham Grindlay, and Daniel PW Ellis, "Transcribing multi-instrument polyphonic music with hierarchical eigen instruments", IEEE Journal of Selected Topics in Signal Processing, Vol. 5, No. 6, pp. 1159-1169, 2011.
[44] Olivier Gillet and Gaël Richard, "Transcription and separation of drum signals from polyphonic music", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 3, pp. 529-540, 2008.
[45] Amelie Anglade, Emmanouil Benetos, Matthias Mauch and Simon Dixon, "Improving music genre classification using automatically induced harmony rules", Journal of New Music Research, Vol. 39, No. 4, pp. 349-361, 2010.
[46] 46. Vishweshwara Rao and Preeti Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 18, No. 8, pp. 2145-2154, 2010.
[47] Vishweshwara Rao, Pradeep Gaddipati, and Preeti Rao, "Signal-driven window-length adaptation for sinusoid detection in polyphonic music", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 1, pp. 342-348, 2012.
[48] Matija Marolt, "Automatic transcription of bell chiming recordings", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 3, pp. 844-853, 2012.
[49] Jia-Min Ren and Jyh-Shing Roger Jang, "Discovering time-constrained sequential patterns for music genre classification", IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 4, pp. 1134-1144, 2012.
[50] Yizhar Lavner and Dima Ruinskiy, "A decision-tree-based algorithm for speech/music classification and segmentation", EURASIP Journal on Audio, Speech, and Music Processing, vol. 2, 2009.
[51] Jingyi Shen, Runqi Wang, Han-Wei Shen, "Visual exploration of latent space for traditional Chinese music", Visual Informatics, vol. 4, no. 2, pp.99-108, June 2020.
[52] Dhara, P., Rao, K.S. “Automatic note transcription system for Hindustani classical music”. Int J Speech Technol, vol. 21, pp. 987–1003, 2018. https://doi.org/10.1007/s10772-018-9554-1.
[53] Akbari, M., Liang, J. & Cheng, H. “A real-time system for online learning-based visual transcription of piano music”. Multimed Tools Appl, vol. 77, pp. 25513–25535, 2018. https://doi.org/10.1007/s11042-018-5803-1.
[54] Bhalarao, R., Raval, M. Automated tabla syllable transcription using image processing techniques. Multimed Tools Appl, 2020. https://doi.org/10.1007/s11042-020-09417-0.
[55] 55. Rodríguez-Serrano, F.J., Carabias-Orti, J.J., Vera-Candeas, P., “Monophonic constrained non-negative sparse coding using instrument models for audio separation and transcription of monophonic source-based polyphonic mixtures”, Multimed Tools Appl, vol. 72, pp. 925–949, 2014. https://doi.org/10.1007/s11042-013-1398-8
[56] Shen, H., Lee, C. An interactive Whistle-to-Music composing system based on transcription, variation and chords generation. Multimed Tools Appl, vol. 53, pp. 253–269, 2011. https://doi.org/10.1007/s11042-010-0510-6.

[57] V. Arora and L. Behera, "Multiple F0 Estimation and Source Clustering of Polyphonic Music Audio Using PLCA and HMRFs," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 2, pp. 278-287, Feb. 2015. doi: 10.1109/TASLP.2014.2387388.

[58] F.J. Cañadas Quesada, N. Ruiz Reyes, P. Vera Candeas, J.J. Carabias & S. Maldonado () A Multiple-F0 Estimation Approach Based on Gaussian Spectral Modelling for Polyphonic Music Transcription, Journal of New Music Research, vol. vol. 39, no. 1, pp. 93-107, 2010. DOI: 10.1080/09298211003695579.

[59] Dorian Cazau, Yuancheng Wang, Marc Chemillier & Olivier Adam, "An automatic music transcription system dedicated to the repertoires of the marovany zither", Journal of New Music Research, vol. 45, no. 4, pp. 343-360, 2016. DOI: 10.1080/09298215.2016.1233285.

[60] Jose J. Valero-Mas, Emmanouil Benetos & José M. Iñesta, "A supervised classification approach for note tracking in polyphonic piano transcription", Journal of New Music Research, vol. 47, no. 3, pp. 249-263, 2018. DOI: 10.1080/09298215.2018.1451546

[61] Jiayin Sun & Hongyan Wang, "A Cognitive Method for Musicology Based Melody Transcription", International Journal of Computational Intelligence Systems, vol. 8, no. 6, pp. 1165-1177, 2015. DOI: 10.1080/18756891.2015.1113749

[62] Tiago Fernandes Tavares, Jayme Garcia Arnal Barbedo & Romis Attux, "Unsupervised note activity detection in NMF-based automatic transcription of piano music", Journal of New Music Research, vol. 45, no. 2, pp. 118-123, 2016. DOI: 10.1080/09298215.2016.1177552

[63] İsmail Arı, Umut Şimşekli, Ali Taylan Cemgil & Lale Akarun (2014) Randomized Matrix Decompositions and Exemplar Selection in Large Dictionaries for Polyphonic Piano Transcription, Journal of New Music Research, 43:3, 255-265. DOI: 10.1080/09298215.2014.891628

[64] Willie Krige, Theo Herbst & Thomas Niesler (2008) Explicit Transition Modelling for Automatic Singing Transcription, Journal of New Music Research, 37:4, 311-324. DOI: 10.1080/09298210902890299.

[65] Paulus, J., Klapuri, A, "Drum Sound Detection in Polyphonic Music with Hidden Markov Models", J AUDIO SPEECH MUSIC PROC. 2009, Article number: 497292, 14 December 2009. https://doi.org/10.1155/2009/497292.

[66] Gao, X., Gupta, C. and Li, H, "Automatic lyrics transcription of polyphonic music with lyrics-chord multi-task learning", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 30, pp.2280-2294, 2022.

[67] Wang, X., Tian, B., Yang, W., Xu, W. and Cheng, W., "MusicYOLO: A Vision-Based Framework for Automatic Singing Transcription", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 31, pp.229-241, 2022.

[68] Wang, Y., Jing, Y., Wei, W., Cazau, D., Adam, O. and Wang, Q., "PipaSet and TEAS: A Multimodal Dataset and Annotation Platform for Automatic Music Transcription and Expressive Analysis Dedicated to Chinese Traditional Plucked String Instrument Pipa", IEEE Access, Vol. 10, pp.113850-113864, 2022.

[69] Simonetta, F., Avanzini, F. and Ntalampiras, S., "A perceptual measure for evaluating the resynthesis of automatic music transcriptions", Multimedia Tools and Applications, Vol. 81(22), pp.32371-32391, 2022.

[70] Wu, H., Marmoret, A. and Cohen, J.E., "Semi-Supervised Convolutive Nmf For Automatic Piano Transcription", 2022.

[71] Holzapfel, A., Benetos, E., Killick, A. and Widdess, R., "Humanities and engineering perspectives on music transcription", Digital Scholarship in the Humanities, Vol. 37(3), pp.747-764, 2022.