applied sciences

MDPI

Review

A Comprehensive Review on Music Transcription

Bhuwan Bhattarai and Joonwhoan Lee *

Center for Advanced Image and Information Technology, Jeonbuk National University, Jeonju 54896, Republic of Korea; bhubon240@gmail.com
* Correspondence: chlee@jbnu.ac.kr

Abstract: Music transcription is the process of transforming recorded sound of musical performances into symbolic representations such as sheet music or MIDI files. Extensive research and development have been carried out in the field of music transcription and technology. This comprehensive review paper surveys the diverse methodologies, techniques, and advancements that have shaped the landscape of music transcription. The paper outlines the significance of music transcription in preserving, analyzing, and disseminating musical compositions across various genres and cultures. It also provides a historical perspective by tracing the evolution of music transcription from traditional manual methods to modern automated approaches. It also highlights the challenges in transcription posed by complex singing techniques, variations in instrumentation, ambiguity in pitch, tempo changes, rhythm, and dynamics. The review also categorizes four different types of transcription techniques, frame-level, note-level, stream-level, and notation-level, discussing their strengths and limitations. It also encompasses the various research domains of music transcription from general melody extraction to vocal melody, note-level monophonic to polyphonic vocal transcription, single-instrument to multi-instrument transcription, and multi-pitch estimation. The survey further covers a broad spectrum of music transcription applications in music production and creation. It also reviews state-of-the-art open-source as well as commercial music transcription tools for pitch estimation, onset and offset detection, general melody detection, and vocal melody detection. In addition, it also encompasses the currently available python libraries that can be used for music transcription. Furthermore, the review highlights the various open-source benchmark datasets for different areas of music transcription. It also provides a wide range of references supporting the historical context, theoretical frameworks, and foundational concepts to help readers understand the background of music transcription and the context of our paper.

Keywords: music transcription; melody transcription; monophonic and polyphonic transcription; deep learning

Citation: Bhattarai, B.; Lee, J. A Comprehensive Review on Music Transcription. *Appl. Sci.* **2023**, *13*, 11882. https://doi.org/10.3390/app132111882

Academic Editor: Lamberto Tronchin

Received: 28 August 2023
Revised: 27 October 2023
Accepted: 29 October 2023
Published: 30 October 2023

1. Introduction

The complex components of music have the power to evoke emotions which can communicate ideas that words alone often struggle to convey. Modern technology has opened up new channels for the expression and preservation of music. One central aspect of this evolution is music transcription, which is the process of converting auditory musical signals into symbolic representations that can be studied, analyzed, and shared. The development of music notation itself is connected to the history of music transcription, which dates back to Mesopotamia and Ancient Greece, where rudimentary notational systems were used to record melodies and rhythms [1–4]. These techniques developed over the years, eventually leading to the complex and standardized music notation that we have today [5–7]. Additionally, music exhibits diversity across various genres and cultures [8,9], each with its unique music notation systems. To accommodate these diversities, different transcription methods are required.

The technological advancements in recent decades have assisted in a new era of music transcription. From the pioneering work of Jean-Jacques Rousseau in the 18th century [10]

to the advent of computers in the 20th century [11], numerous scholars and inventors have contributed to the development of transcription techniques. Today, with the rise of digital signal processing, machine learning, and artificial intelligence [12–17], there is a significant and substantial change occurring in the field of music transcription [18–22].

The aim of this comprehensive review paper is to examine an in-depth exploration of the various methodologies, techniques, and advancements that have influenced the area of music transcription. By synthesizing a wide range of research studies, we seek to offer a holistic understanding of the challenges, opportunities, and future directions in music transcription. To achieve this goal, we have organized the paper as follows.

In Section 2, we provide an overview of music transcription by tracing from frame-level to note-level, then to stream-level, and finally to notation-level transcription. Section 3 delves into the theoretical foundations of vocal melody transcription, highlighting the complexities of monophonic and polyphonic vocals in both frame-level and note-level transcription tasks. In addition to this, we also explore the various insights and advances in music transcription areas like multi-pitch estimation and single-instrument as well as multi-instrument transcription. In Section 4, we list the current challenges of music transcription methodologies into several main factors: singing techniques, instrumentation, pitch ambiguity, tempo changes, and subjectivity. Section 5 discusses various available open-source as well as commercial music transcription tools. Section 6 describes the various python libraries that are useful for music transcription. In Section 7, we explore the applications of music transcription in various fields like music education, music creation, and music production. Evaluating the effectiveness of transcription systems is a critical aspect. So, we review the established benchmark datasets in Section 8 ranging from vocal melody to general melody, from monophonic vocals to polyphonic vocals, and from single-instrumentation to multi-instrumentation. The future directions are explained in Section 9. Finally, the review concludes in Section 10, with a discussion on the broader implications of music transcription.

Throughout this review paper, we draw upon a diverse array of references encompassing music theory, signal processing, machine learning, and human–computer interaction. The contributions of scholars, researchers, and innovators from around the world have collectively shaped the trajectory of music transcription. By synthesizing these contributions, we aspire to offer a comprehensive resource that inspires further advancements and collaborations in this dynamic field.

2. Overview of Music Transcription

This section presents four different levels of music transcription: frame-level transcription, note-level transcription, stream-level transcription, and notation-level transcription. At the most fundamental level, frame-level transcription involves analyzing very short frames of audio and identifying basic sound characteristics of music such as pitch and timbre. In the next step, note-level transcription takes the information gathered from frame-level analysis and assembles it into individual musical notes. This level of transcription can identify the pitch, duration, and timing of each note in the piece of audio. The stream-level transcription looks beyond individual notes to capture the larger musical phrases. This involves recognizing patterns in the sequence of notes, identifying chord progressions, and determining the overall structure of the music like verses and choruses in a song. The highest level of transcription involves creating a formal written representation of the music notation. This includes not only the individual notes and their timing but also various musical symbols such as dynamics, articulations, time signatures, key signatures, and more. These four levels of transcription, frame-level, note-level, stream-level, and notation-level, are illustrated in Figure 1.

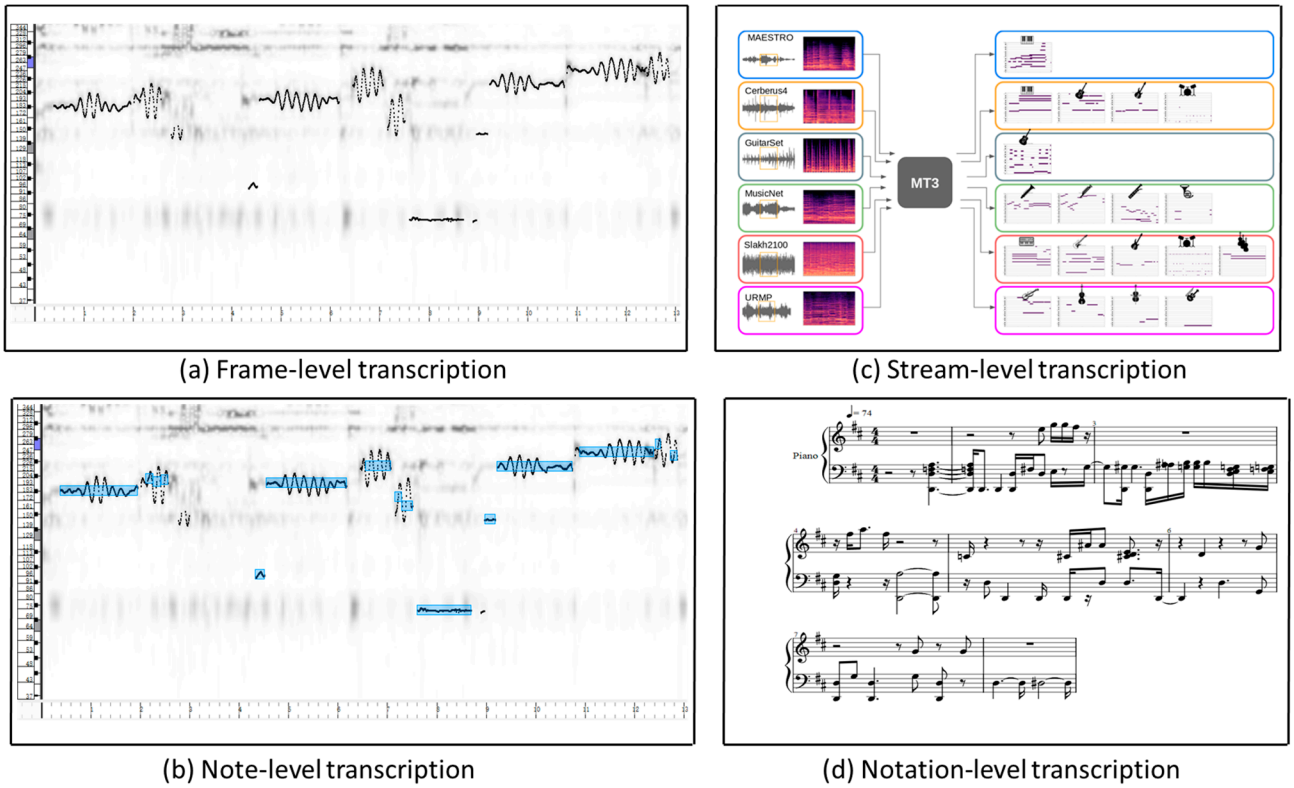

Figure 1. An example for the illustration of four different levels of music transcription.

2.1. Frame-Level Transcription

Frame-level transcription or multi-pitch estimation (MPE) is the estimation of the pitch of notes that are simultaneously present in each time frame of the audio signal. This transcription task has been performed in each time frame independently. Figure 1a shows an example of frame-level transcription, where every dot in the x-axis represents the discrete time and the y-axis dots represent its corresponding pitch. The visualization for this transcription has been carried out using Tony [18] from the wave file (opera_male5.wav) of the ADC2004 (http://labrosa.ee.columbia.edu/projects/melody/, accessed on 5 October 2023) vocal melody dataset. Various research studies have been developed using frame-level transcription. The work in [23] introduced a method for estimating multiple concurrent piano notes, dealing with overlapping overtones by using smooth autoregressive models. It handles background noise to reduce pitch estimation errors, creating a comprehensive model for piano notes. The effectiveness of this model has been validated with real piano recordings of the MAPS [23] dataset. Similarly, the work in [24] discusses the challenge of training data for estimating multiple pitches in music. It introduces a new approach called combined frequency and periodicity (CFP), which combines two different types of features to improve the accuracy in simultaneous pitches. The result shows that CFP works well for western polyphonic music and remains effective even when dealing with audio distortions like filtering and compression. The two works described above are conventional approaches for frame-level transcription. However, recently, the neural network-based methods [12,25] have garnered increased attention from researchers. The work in [12] introduced a supervised neural network model for a polyphonic transcription task on the MAPS dataset. It consists of an acoustic model and a music language model. The acoustic model, based on neural networks, estimates pitch probabilities in audio frames, while the recurrent neural network-based language model captures pitch correlations over time. The acoustic and music language model predictions are combined using a probabilistic graphical model, with inference performed using the beam search algorithm. The experimental results demonstrate that convolutional neural network-based acoustic

models outperform popular unsupervised models. Similarly, the work in [25] analyzed neural network based frame-level piano transcription by comparing four different input representations: spectrograms with linearly spaced bins, spectrograms with logarithmically spaced bins, spectrograms with logarithmically spaced bins and a logarithmically scaled magnitude, as well as the constant-Q transform. This study highlights the significance of choosing the right input representation and fine-tuning the hyper-parameters, especially the learning rate and its schedule, which can improve the accuracy of frame-level transcription. Frame-level transcription has several advantages as well as disadvantages. First, it provides a high temporal resolution which can enable the detailed representations of music at a fine time scale. Second, it is useful for in-depth music analysis where precise timing information is required. However, there are some notable disadvantages in frame-level transcription. First, it requires high computational resources due to the need to analyze audio at a very fine level, which can make it less efficient for real-time applications. Second, the fine-scale high temporal resolution generates a large volume of data, which makes it challenging to manage, store, and process efficiently due to its size and complexity. Lastly, it focuses deeply on the technical details but fails to capture the broader musical context.

2.2. Note-Level Transcription

The note-based transcription approaches directly estimates notes, including pitches and onsets. This note-based transcription is one level higher than frame-level transcription because the objective of frame-level transcription is to estimate only pitch in a particular time period. But note-level transcription needs to estimate the pitch, onset time, and offset time of a specific note in a particular time period. Figure 1b shows an example of note-level transcription. The blue rectangle shows the note events after the post-processing of frame-level transcription on a continuous pitch based on a hidden Markov model. The visualization for this transcription has also been carried out using Tony [18] from the wave file (opera_male5.wav) of the ADC2004 vocal melody dataset. The note offset in this level of transcription is ambiguous, so they are sometimes neglected during the inference time and only evaluate the pitch and onset time of a note. The work in [26,27] estimates the pitches and onsets in a single architecture. One example for this is [28], which jointly estimates the different attributes of notes like the pitch, intensity, onset, and duration. It estimates the properties of notes concurrently using harmonic temporal structured clustering. The note-level transcription can also be achieved after the post-processing in frame-level transcription. First, the fundamental frequency is estimated concurrently, and post-processing is applied to estimate the musical notes in the second step. The methods used during the post-processing steps are median filtering, hidden Markov Models, and Neural networks [24,29,30]. The work in [24] uses a median filter by comparing the estimated pitch in nearby frames for temporal smoothing. The moving median filter is used with 0.25 s, with a hop size of 0.01 s. This post-processing method is reliable for connecting non-continuous pitch and can effectively delete the isolated one. Similarly, the work in [29] converted the output of the support vector machine to a posterior probability. The steps for pitch smoothing were performed for each note class by using the Viterbi search method on the transition matrix of 2×2. The note onset and offset were finally gathered from both the posterior probability of the support vector machine and the training data. Note-level transcription can provide the highest level of detail, capturing individual notes' onset and offset along with pitch. It allows for an in-depth analysis of musical elements like melody, harmony, and rhythm. However, there are some disadvantages of note-level transcription. First, transcribing note by note can be time-consuming, which is not practical for longer pieces of music. Second, to make an accurate transcription, expert musicians are required for the annotation of onset, offset, and pitch. Third, it is challenging and difficult to capture the musical nuances like vibrato and articulation in note level-transcription because it is highly subjective and depends on the interpretation and style of the performer.

2.3. Stream-Level Transcription

Stream-level transcription is also called multi-pitch streaming (MPS), which can group estimated pitches or notes into separate streams. The grouped or separated stream typically corresponds to an individual instrument. Figure 1c shows an example of stream-level transcription which was obtained from a multi-task multitrack music transcription (MT3) model [31]. The estimated pitches and notes for each instrument in this model have been grouped into separate streams using various music transcription datasets. The pitch streams of different instruments in this transcription can be visually distinguished using varying colors. This level of transcription offers a broader perspective, involving the transcription of musical events or attributes across entire segments or streams of music. These segments may span longer durations and encompass multiple notes or musical phrases. Stream-level transcription has the capacity to capture higher-level musical structures, such as chords, sections, or phrases, thereby aiming to provide a comprehensive and holistic representation of the analyzed music. This level of transcription is useful when the goal is the simplified arrangement of a musical piece. It works well for music with a smaller number of notes. However, it reduced some level of detail compared to note-level transcription, which may limit the accuracy of the transcription. Stream-level transcription is less suitable for complex music because it captures larger musical sections, which means it provides a more generalized overview of the music but does not capture the complex compositions such as rapid note sequences or variations in dynamics. In stream-level transcription, recent research efforts [32] have addressed the challenge by considering three key inputs: the original audio mixture, an existing multi-pitch estimation system, and the number of sources within the audio mixture. The approach undertaken in this work formulates the transcription problem as a clustering task, with a focus on maintaining timbral consistency. This is achieved by assuming that sound objects originating from the same source exhibit similar timbral characteristics. Similarly, other studies [33,34] serve as additional examples of stream-level transcription. Both of these works also operate under the assumption that notes belonging to the same source stream demonstrate comparable timbral traits in contrast to notes from different source streams.

2.4. Notation-Level Transcription

Notation-level transcription is the higher level of transcription which is the process of converting audio recordings into a traditional musical notation format such as sheet music or a musical score. Figure 1 illustrates an example of notation-level transcription. The visualization for this transcription has been carried out using AnthemScore [35] from a wave file (opera_male5.wav) of the ADC2004 vocal melody dataset. This transcription maps the audio signals to specific symbols, notes, rhythms, and other notation elements that can be read and performed by musicians. It focuses on capturing the basic structure, melody, and harmony, making it suitable for less complex music. It provides a quick overview of the musical content which can be easily understood by individuals with limited musical training. However, it is not suitable for in-depth musical analysis or the complex composition of the music. Notation-level transcription is often used to preserve and communicate musical compositions and performances in a standardized form. The transcription in this level requires a deeper understanding of musical structures, including harmonic, rhythmic, and stream structures. Transcribers use specialized music notation software to create the symbolic representation of the music. The popular software for notation-level transcription include Finale V27 (https://www.finalemusic.com/, accessed on 5 October 2023), Sibelius for desktop and mobile version 2023.8 (https://www.avid.com/sibelius, accessed on 5 October 2023), MuseScore 4.x (https://musescore.org/, accessed on 5 October 2023), LilyPond 2.24.2 (http://lilypond.org/, accessed on 5 October 2023), AnthemScore version 4 [35], and more. These tools provide a user-friendly interface for inputting and editing musical elements.

3. Transcription Techniques: Insights and Advances

This section explores a range of transcription techniques beginning with general melody to vocal melody extraction for extracting dominant melodies from polyphonic audio. The dominant melody could be either vocal or any other instruments. It then explores multi-pitch estimation that can concurrently estimate the pitch of different components of audio simultaneously. Note-level transcription for solo and ensemble vocals has also been carried out by showing the techniques for capturing the uniqueness of vocal expressions. The section ends up with a discussion of single-instrument and multi-instrument transcriptions, explaining the complex connections between various sound sources and harmonics in musical recordings.

3.1. General Melody Extraction

General melody extraction is the sub-task of frame-level transcription that deals with identifying the pitch information of the dominant melody present in the polyphonic audio. The research work in [36] proposed a new encoder/decoder network for extracting melodies from polyphonic musical audio. The authors offer two key contributions: first, they use the semantic pixel-wise segmentation method to improve melody localization in frequencies with fewer convolutional layers. Second, they propose utilizing the network's bottleneck layer to estimate the presence of a melody line for each time frame in the audio. In general melody extraction, each frame is analyzed to estimate the fundamental frequency or pitch of the most prominent musical note. The early work presented in [37] extracts the melody played by a solo instrument by generating multiple fundamental frequency (F0) hypotheses at each time frame and employing various knowledge sources, including instrument recognition and temporal information, to select the most likely F0 sequences. Additionally, the work in [38] introduced a fully convolutional neural network trained on a large, semi-automatically generated dataset for fundamental frequency estimation. The model proves the effectiveness in both multi-fundamental frequency (multi-F0) and melody tracking in polyphonic audio, achieving state-of-the-art results across various datasets. The application of general melody extraction is often used in instrumental music analysis and can help in tasks such as automatic transcription, cover song generation, and music recommendation [39,40].

3.2. Vocal Melody Extraction

Vocal melody extraction is the sub-task of frame-level transcription. It is the process of identifying the sung or spoken pitch in a vocal performance [36,41,42]. The encoder/decoder network described in Section 3.1 for general melody extraction [39] is also used for vocal melody transcription. The work in [41] uses a high-resolution network (HRNet) to separate vocals from polyphonic music and an encoder–decoder network to estimate vocal F0 values. The experimental result shows that the HRNet-based singing voice separation method effectively reduces the accompaniment interference, which outperforms other state-of-the-art algorithms in most cases. Similarly, the work in [42] introduces a novel melody extraction system that treats the problem as semantic segmentation on a time-frequency image. It distinguishes melody from non-melody elements by using a deep convolutional neural network (DCNN) with dilated convolution. The utilization of an adaptive progressive neural network which transfers the knowledge from symbolic to audio data demonstratea the system's accuracy across different datasets. Similar to general melody extraction, vocal melody involves F0 estimation for each frame, but with a focus on capturing the melody carried by the human voice [43–45]. Inspired by the object detection model in image processing, the research work in [43] introduced a patch-based CNN model for extracting vocal melodies. This patch-based CNN allows for efficient training with limited labeled data. Similarly, the work in [44] utilized the power of semi-supervised learning using a teacher–student network, which outperforms supervised learning in terms of accuracy. The work in [45] uses multi-task learning to train the classification and detection network for singing voice detection and pitch estimation at the same time. The vocal

melody extraction is essential for applications like lyrics-to-audio synchronization, karaoke generation, and vocal synthesis.

3.3. Multi-Pitch Estimation

Multi-pitch estimation or multi-F0 estimation is a signal processing and music analysis technique used to identify and track multiple simultaneous pitches in an audio signal [34,46–50]. Many works have been carried out using this approach. The work in [34] clusters the multi-pitch F0 values into different sources by using probabilistic latent component analysis (PLCA). Similarly, the work in [46] outlines a system developed for the MIREX 2009 contest in Multiple Fundamental Frequency Estimation and Tracking, utilizing frame-by-frame analysis and a tracking mechanism. Additionally, the computational model in [47] involves the human auditory periphery followed by a novel periodicity analysis. This method is computationally efficient and straightforward to implement since the peripheral hearing model needs to be computed only once. Moreover, the proposed algorithm in [48] operates at frame-level for multi-F0 estimation, searching for the set of fundamental frequencies that minimize a spectral distance measure in each audio frame. The spectral distance measure is defined under the assumption that a polyphonic sound can be modeled by a weighted sum of Gaussian spectral models. Similarly, the book and tutorials in [49,50] outline the fundamental concepts, statistical methods, filtering methods, and subspace methods related to pitch estimation, making it a valuable resource for those interested in this field. Multi-pitch estimation is particularly important in scenarios where multiple sound sources, such as different musical instruments or vocalists, are present together in the same recording [51]. Multi-pitch estimation has applications in various fields, including music transcription, source separation, music analysis, and audio effects processing. In music transcription, it can be used to extract the individual notes played by different instruments in a polyphonic musical piece. In speech processing, it can help separate and analyze multiple speakers in an audio recording.

3.4. Solo Vocal Note Transcription

The note-level solo vocal transcription focuses specifically on capturing the sung or spoken notes within a solo vocal performance in a song [22,52,53]. The work in [22] introduces VOCANO for singing note transcription (SVT) using multi-task semi-supervised learning techniques. The work transcribes the vocal notes of singing from polyphonic audio using virtual adversarial training (VAT). Similarly, the probabilistic transcription approach is carried out in [52] to solve the problems of pitch fluctuations in singing voice transcription. It employs a hierarchical Hidden Markov Model (HMM) to manage note transitions and pitch variations, achieving promising results when assessed on a monophonic sung melody dataset. A huge number of labelled data is necessary to train the supervised deep learning model. Keeping this in mind, the work in [53] introduces a novel contrastive learning-based data cleansing model after identifying the incorrect labels in vocal note event annotations. The transcription process for solo vocals is more complex than in instrumental music due to the inherent variability and expressiveness of the human voice [54]. Note-level transcription for vocals involves accurately determining the pitches, timings, and durations of each sung note [18,55]. To correctly identify vocal notes, the work in [18] uses Viterbi decoding as a post-processing technique after following pitch estimation in monophonic audio. To overcome the limited number of labeled training data, the work in [55] constructs a dataset of solo vocals that includes the annotations in frame-level f0, notes, lyrics, and track-level meta-data. To accurately transcribe the notes of solo vocals, the machine learning-based models must consider factors like vibrato, ornamentation, and changes in articulation that contribute to the unique character of the vocal performance. This type of solo vocal note transcription is essential for applications like lyrics synchronization, vocal synthesis, and cover song generation.

3.5. Vocal Ensemble Transcription

Note level transcription in vocal ensembles is a challenging task because vocal lines are simultaneously occurring in an audio recording [56–58]. The comprehensive review of choir acoustics provided in [56] is particularly valuable for future researchers engaged in vocal transcription studies. The research in [57] investigates the intonation accuracy in four-part of singing (SATB) by comparing individual singers and collaborative ensembles, examining how different listening conditions affect pitch accuracy. Similarly, the work in [58] introduces a spectrogram factorization method for the automatic transcription of a cappella performances with multiple singers. It employs a sparse dictionary of spectral templates for vowel vocalizations and includes a post-processing step using overtone-based features to reduce false-positive in pitch detections. The process in vocal ensemble requires identifying and separating the various vocal lines, estimating the pitches and timings of each line's individual notes, and creating a comprehensive representation of the entire vocal ensemble [51,59]. The early work in [59] investigated the preferences of experienced listeners for pitch and formant frequency dispersion in unison choir sounds using synthesized stimuli. The later work in [51] utilizes convolutional neural networks for extracting F0 values in ensemble singing. The method is trained on a dataset containing vocal quartets with F0 annotations and is evaluated across various scenarios, including recordings with reverb. Transcription in vocal ensemble is essential for analyzing and understanding multi-part vocal arrangements, harmonies, and interactions between different voices. It has applications in music analysis, arrangement, and synthesis, enabling the recreation of complex vocal textures from recorded performances. The vocal ensemble transcription can distinguish between different vocal ranges—soprano, alto, tenor, and bass—which requires special analysis to extract the individual contributions of each voice [60]. This process enables the reconstruction of multi-part vocal arrangements and the interpretation of polyphonic vocal music [61].

3.6. Single-Instrument Transcription

Single-instrument transcription is the process of converting an audio recording containing a single musical instrument into a symbolic representation, such as musical notes or sheet music. Most of the previous automatic music transcription focused on single-instrument piano transcription [25,62–65]. The proposed methods in [25,62,63,65] use deep learning-based approaches for piano transcription. Similarly, the proposed method in [64] uses non-negative matrix factorization by modeling a piano note as a percussive and harmonic decay in a supervised way. The work in [25] is designed only for frame-level transcription, while the work in [62] jointly estimates the onset events and pitch. Similarly, the method proposed in [63] can estimate the onset and offset times with an arbitrary time resolution, which outperforms the accuracy in the MAESTRO dataset [66]. Similarly, the work in [65] automatically predicts the piano note onset events given a video of a person playing a piano. The works described above are all for piano transcription. However, there are some other single-instrument transcriptions like guitar [67,68], violin [69,70], and flute transcriptions [71,72]. The work in [67] is the conventional approach which iteratively processes frequency peaks by marking fundamental frequency, while [68] is the recently published paper based on an attention mechanism and convolutional neural network for guitar transcription. Similarly, the research works in [69,70] discuss the analysis of violin music, but they have different objectives and methods. Additionally, [69] focuses on transcribing the pitch, duration, and playing techniques based on the time and frequency domain properties of the audio signal. In contrast, [70] aims to recover fingerings from violin music to recreate timbre, using the hidden Markov model. Moreover, the work in [71,72] is related to the transcription of Arabian flute music using audio features like Discrete Fourier Transform (DFT) and Fast Fourier Transform (FFT). The work in [71] is designed to detect octave falls and to perform the mapping from intensity to MIDI velocity, while [72] detected the Arabian flute pitches, which improved the accuracy over the occidental flute.

3.7. Multi-Instrument Transcription

Multi-instrument transcription involves transcribing an audio recording containing multiple musical instruments or sound sources. The objective is to separate and extract the individual musical lines of each instrument, capturing their respective pitches, timings, and contributions to the overall musical texture. Recently, deep learning-based multi-instrument transcription [73–77] has been widely used and is crucial for analyzing and understanding the interactions between different instruments in a polyphonic musical piece. A recently published paper [73] jointly considers the instrument recognition module, the transcription module, and the source separation module, which is capable of transcribing, recognizing, and separating multiple musical instruments from the audio signal. Similarly, the work in [74] adapts the concept of computer vision methods like multi-object detection and instance segmentation for multi-instrument note tracking. The multi-instrument multi-pitch estimation (MI-MPE) task, which is a clustering-based method for music transcription, was proposed by [75]. It estimates piano rolls for diverse musical instrument categories. The authors in this article claim that their method became the very first MI-MPE to be published. Similarly, the paper in [76] implements the concept of multi-object semantic segmentation to solve the problem of pitch activity detection and instrument recognition. Moreover, the work presented in [77] focuses on transcribing snare drums, bass drums, and hi-hats using a convolutional neural network and convolutional recurrent neural network after synthesizing a large number of training data. There are some other methods that carry out multi-instrument transcription after source separation. In this type of multi-instrument transcription task, the algorithm first separates each independent instrument using a source separation algorithm and estimates the pitch and note for each source separately [78,79]. Finally, the transcribed notes and musical elements from each instrument are combined to create a comprehensive representation of the polyphonic music, capturing the interplay between different instrumental lines.

4. Challenges for Music Transcription

Polyphonic music is made up of multiple simultaneous sources that include a variety of instruments and vocals with varying pitches, volumes, and timbres, each of which produces one or more musical components. The difficulty of inferring musical qualities from a mixing signal is exceedingly understudied. Moreover, vocal transcription is currently facing a primary challenge in the field of music transcription due to the inherent complexity and diversity of singing techniques employed by vocalists [80,81]. Unlike many musical instruments, the human voice is incredibly versatile and capable of producing a wide range of timbres, dynamics, articulations, and expressive nuances. This inherent variability gives rise to several factors that make vocal transcription more challenging. In addition to this, the harmonic relationship between overlapping sound events in music makes it complex for any type of transcription. The annotation of ground-truth for polyphonic music takes a long time and demands a lot of skill. This creates a limitation for the powerful supervised machine learning techniques. The following points describe the challenges for music transcription in more detail.

4.1. Singing Techniques

Vocalists can produce an extensive array of timbres and tonal qualities, ranging from breathy and soft to powerful and resonant [82,83]. These timbral variations are influenced by factors such as vocal tract shaping, tension, and breath control [84]. Accurately capturing and representing these timbral changes in vocal transcription requires sophisticated analysis techniques. Singers often employ glissandi (continuous pitch slides) and microtonal inflections, which makes vocal transcription challenging in precisely determining discrete pitch values [85,86]. The continuous nature of vocal pitch transitions requires advanced algorithms to track and transcribe these pitch changes accurately.

4.2. Instrumentation

The type of instrument used in a recording can significantly affect the transcription process due to its distinct characteristics, timbres, and playing techniques associated with different instruments [87]. For example, transcribing piano music involves handling multiple simultaneous notes across a wide range [88,89]. Each musical instrument consists of distinctive timbres, which arise from the combination of the physical attributes and material and playing techniques of singers. Transcribing instruments with diverse timbral characteristics requires advanced algorithms to accurately capture and improve the performance of music transcription. In addition to this, instruments have varying pitch ranges which influence the frequency range of the audio signal. The handling for this problem requires more sophisticated transcription algorithms that must be effective in tracking pitch variations across the spectrum.

4.3. Pitch Ambiguity

Some audio segments may have multiple possible interpretations, leading to ambiguity in choosing the correct pitch or note. This phenomenon can occur due to various reasons, such as harmonics, overtones, noise, or the limitations of pitch estimation algorithms [90,91]. Pitch ambiguity poses a significant challenge for accurate music transcription because it can lead to incorrect or uncertain note identifications. Musical sounds often consist of a fundamental frequency and its harmonics. The presence of strong harmonics can lead to ambiguity in pitch estimation, as the algorithm may detect multiple potential fundamental frequencies that align with different harmonics.

4.4. Tempo Changes

Tempo changes, also known as tempo fluctuations or tempo variations, refer to shifts in the speed or pacing of a musical performance. These changes occur when the tempo, or the perceived speed of the music, deviates from a consistent or steady rhythm [92]. Variations in tempo can affect the perceived durations of notes [93,94]. Faster tempi might result in shorter note durations, while slower tempi can lead to longer note durations. Sudden tempo changes can make the rhythm ambiguous, making it challenging to accurately transcribe the timing and alignment of notes. Tempo changes can disrupt the perception of the underlying beat or pulse, making it difficult to identify beats and downbeats properly.

4.5. Subjectivity

Music transcription involves the process of converting an audio recording of music into a symbolic representation, such as sheet music or a score [95,96]. While transcription aims to capture the essential elements of the music, it is important to acknowledge that there is an inherent degree of subjectivity involved in the process. This subjectivity arises from the fact that different transcribers, even with expertise in music theory and transcription techniques, may interpret the same musical performance in slightly different ways [94].

5. Music Transcription Tools

The process of translating audio recordings into some symbolic notation has been greatly facilitated by the scope of specialized software tools. These tools offer various transcription needs by accommodating both monophonic and polyphonic musical contexts. The review of well-known tools in this section highlights their features and contributions to the field of music transcription.

Tony [18] stands as a prominent choice among the tools made to handle monophonic music. It can implement both melody estimation methods and note tracking based on pYIN (probabilistic YIN) [19], which is an extension of the YIN algorithm [20], originally developed for pitch detection. Another popular tool called IntelliScore (http://www.intelliscore.net/, accessed on 5 October 2023) of polyphonic version is particularly known for its capability to transcribe polyphonic audio, which means it can analyze and separate multiple instruments or voices that are playing simultaneously in a recording. It can

directly convert audio recordings into an MIDI format that can represent musical notes, rhythms, and other forms of musical elements. The key features of IntelliScore include batch processing for polyphonic transcription, instrument recognition, note detection, and tempo estimation. The users are able to listen to the transcribed MIDI output and make corrections and adjustments based on what they determine to be incorrect. IntelliScore can be used for various purposes including transcribing recorded music, creating MIDI tracks for remixes or arrangements, and extracting melodies or instrumental parts from recordings.

Among the music transcription tools described above, there are many other popular tools available online. Amazing Slow Downer for windows, Mac, and iOS (https://www.ronimusic.com, accessed on 5 October 2023) provides MIDI software such as sweet MIDI player pro, sweet sixteen MIDI sequencer, and sweet MIDI converter. AnthemScore version 4 [35] utilizes artificial intelligence to automatically transcribe complex compositions. AnthemScore can analyze an audio file, such as an MP3 or WAV, and convert it into sheet music or an MIDI file. This can be particularly useful for musicians who want to learn to play a piece of music, composers who want to transcribe their own compositions, or educators who need to notate music for teaching purposes. Sibelius for desktop and mobile version 2023.8 (https://www.avid.com/sibelius, accessed on 5 October 2023), is a powerful notation software for precise musical representation. It provides a comprehensive set of tools for creating and editing music notation. Users can input notes, rests, dynamics, articulations, lyrics, chord symbols, and other musical symbols using a graphical interface. Sibelius supports the MusicXML file format, which allows users to import and export scores between different music notation software. This makes it easier to collaborate with musicians who use different notation software. ScoreCloud version 4.7 [97] employs audio recognition technology for intuitive melody-to-sheet music conversion. Users can sing or play an instrument into a microphone connected to their computer, and ScoreCloud will attempt to notate the music in real-time. Capo version 4 (http://supermegaultragroovy.com/products/capo/, accessed on 5 October 2023) isolates and analyzes audio sections for detailed transcription. It mainly focuses on music transcription, chord detection, and pitch and tempo adjustment. Capo provides visual representations of chords, waveforms, and spectrograms from which the users can better understand the music. Audacity version 3.2 [98] provides basic transcription like beat detection, pitch shifting, and time stretching.

6. Python Libraries for Music Transcription

There are various python libraries available on the internet that can collectively offer a range of capabilities, from comprehensive score analysis to MIDI manipulation, deep learning-based transcription, specialized pianoroll processing, and advanced audio analysis. Their unique strengths provide various aspects of music transcription for researchers, developers, and musicians with diverse tools for musical compositions.

Librosa [99] is a prominent library renowned in music and audio analysis. The primary focus of Librosa lies in audio processing, not in transcription, but it plays a crucial role in transcription-related workflows. It supplies a python build-in function for estimating beat, pitch, and tempo. By extracting this information from audio recordings, Librosa serves as an essential preliminary step for accurate and insightful music transcription. Another popular music transcription tool is Music21 [100]. It is a comprehensive toolkit designed to facilitate tasks in computer-aided musicology. With its rich feature set, it enables the parsing, manipulation, and analysis of musical scores. Music21 supports multiple music notation formats, including MIDI and MusicXML, making it a versatile choice for transcribing and processing music data. Researchers and musicians alike benefit from its capabilities for extracting musical features and converting between different notation representations. The next python library for music transcription is Midiutil 1.2.1 (https://github.com/MarkCWirt/MIDIUtil, accessed on 5 October 2023). It is a library specifically designed for creating and manipulating MIDI files. While it does not provide direct music transcription capabilities, it is useful for generating MIDI files

from scratch or modifying existing ones. This can be handy for tasks such as converting transcribed music data into an MIDI format. Similarly, the transcription library called Pypianoroll [21] was designed especially to handle multitrack pianorolls data, which is a compact representation of musical scores often used in electronic music. While not strictly for music transcription, it can be useful for tasks involving pianoroll-based music analysis, such as generating and modifying MIDI-like sequences. Moreover, another popular python library called madmom [101] focuses on research and experimentation in the field of music information retrieval. It is suitable for both novice users who want to perform basic transcription tasks and researchers who want to develop and test new algorithms for more advanced applications. It provides a wide range of tools and functions for analyzing audio signals and extracting useful information from them. Music transcription in madmom involves converting audio recordings of music into a symbolic representation, often in the form of musical notes, chords, or other musical elements. Some other popular python libraries related to transcription include Essentia 2.0.1, (https://github.com/MTG/essentia, accessed on 5 October 2023) [102], Aubio 0.4.9, (https://github.com/aubio/aubio, accessed on 5 October 2023) [103], and pydub 0.25.1 (https://github.com/jiaaro/pydub, accessed on 5 October 2023).

7. Applications of Music Transcription

A successful automatic music transcription (AMT) system would allow for a wide range of interactions between people and music. It is an enabling technology with clear potential for both economic and societal impacts. The most well-known application of automatic music transcription is enabling musicians to capture the notes of an improvisational performance so that they can replicate it afterwards. AMT is also useful for generating musical styles in the absence of a score. Due to the numerous applications associated with the area, such as the automatic search and annotation of musical information, interactive music systems (e.g., computer participation in live human performances, score following, and rhythm tracking), and musicological analysis, the problem of automatic music transcription has attracted considerable research interest in recent years. The various applications for AMT are highlighted below:

Music education: It can be used to make a system that can help for the automatic tutoring of various types of instruments [104,105].

Music creation: It can be used to improve and create the music by automatically identifying the musical notes and accompaniment [106,107].

Music production: It can be used for music content visualization and intelligent content-based editing [108].

Music search: It can be used for indexing the various genres of music such as pop, rock, metal, jazz, etc. It can also be used for recommendation of various musical lines such as the melody, chord progression, bass, and rhythm [109].

Musicology: It can be used for analyzing jazz improvisations and unannotated music [110].

Vocal synthesis: Transcribing vocals is crucial for vocal synthesis and harmonization technologies. It allows for the creation of realistic vocal performances or harmonies through software and synthesizers [111,112].

Music recommendation system: The data transcribed by an AMT system can be used for music recommendation systems as tools for suggesting relevant music to the users by analyzing musical trends and listening habits [113].

Karaoke generation: Music transcription contributes to karaoke systems allowing users to sing along accurately to their favorite songs [114,115].

Music research: The transcribed examples using AMT can support research scholars in analyzing the different topics of music including melody structures, rhythmic patterns, and chord and harmonic progressions [116].

Among the applications described above, AMT systems can also be used for cover song creation and detection, music archiving and restoration, music copyright and licensing, music arrangement and orchestration, and many more music analysis-related tasks.

8. Music Transcription Datasets

Music transcription datasets play a crucial role in transcription research. So, in this section, we will provide a comprehensive overview of these datasets. The datasets we will describe encompass various aspects, including vocal melodies, general melodies, polyphonic vocals, multi-pitch estimation, single-instrumentation, and multi-instrumentation. A detailed description of these datasets is shown in Table 1 below.

Table 1. The description of currently available music transcription datasets.

Dataset Name	Propose	Description
MIR-1K [117]	Vocal Melody	The MIR-1K dataset offers a diverse collection of music excerpts spanning various genres, with detailed annotations of pitch contours in semitones, indices, and types of unvoiced frames, lyrics, and vocal/non-vocal segments, making it an ideal resource for vocal transcription research. It contains 1000 song clips in which the musical accompaniment and singing voice are recorded at the left and right channels.
VOCADITO [54]	Vocal Melody	VOCADITO offers four different types of annotation: frame-level f0, notes, lyrics, and track-level meta-data. It consists of 40 short segments of monophonic singing in seven different languages. First, the frame-level and note-level annotations were constructed using the Tony [13] algorithm. It also allows the user to manually correct mistakes made by the algorithm.
Medley-DB [118]	Vocal/General Melody/Multi-instrument	Medley-DB, originally designed for multi-instrumental analysis, also includes isolated vocal stems that enable researchers to isolate and transcribe vocal melodies accurately. It can not only be used for a multi-instrumental context but also allows users to transcribe general melodies. There are two versions of the MedleyDB dataset. The first version of MedleyDB consists of 122 multitracks of mix, processed stems, raw audio, and metadata, while the second version of MedleyDB 2.0 consists of 74 new tracks, making a total of 196 multitracks.
MIREX05 (https://nema.lis.illinois.edu/nema_out/mirex2013/results/ame/mrx05/, accessed on 5 October 2023)	Vocal Melody	The MIREX05 dataset, utilized in the Music Information Retrieval Evaluation eXchange, features pop songs with annotations for vocals, providing a valuable platform for vocal melody extraction. There are a total of 13 wav files along with their corresponding pitch in the MIREX05 dataset.
ADC2004	Vocal Melody	The ADC2004 dataset, from the Audio Description Contest 2004, offers audio tracks with descriptions of different musical aspects, including vocals, facilitating research into vocal transcription techniques. There are a total of 21 wav files along with their corresponding pitch annotation.
Bach10 [119]	Polyphonic vocal/ multi-pitch estimation	The Bach10 dataset features a collection of compositions by J.S. Bach, encompassing both vocal and instrumental melodies. The dataset comprises audio recordings for each individual part and the complete ensemble of ten four-part chorales composed by J.S. Bach. Additionally, it includes MIDI scores, precise alignment data between the audio and scores, as well as accurate pitch values for each part and note information for each composition. The audio recording of the Soprano, Alto, Tenor, and Bass of each piece is performed by a violin, clarinet, saxophone, and bassoon, respectively.

Table 1. *Cont.*

Dataset Name	Propose	Description
TONAS [120]	Polyphonic vocal	TONAS offers a unique collection of 72 Flamenco songs annotated in semitones, presenting a diverse range of vocal styles and melodic structures. It consists of three cappella singing styles, Deblas, and two variants of Martinete.
Bach chorales [121]	Polyphonic vocal	The Bach chorales dataset, comprising harmonized vocal compositions, is widely used for studying both melodic and harmonic elements in polyphonic vocal music.
MAPS [23]	Single-instrument (Piano)	The MIDI aligned piano sounds (MAPS) dataset centers on solo piano recordings, making it an excellent resource for piano transcription tasks. It is composed of isolated notes, random-pitch chords, and usual western chords. The corrected version of the MAPS dataset has been created by screening the error using an algorithmic strategy [122]. MAPS contains 40 GB of music, which represents about 65 h of recordings.
MAESTRO [66]	Single-instrument (Piano)	MAESTRO offers a collection of classical piano performances, enriching the dataset's diversity and enabling researchers to explore nuances in solo piano transcription techniques. The dataset is composed of about 200 h of paired audio and MIDI recordings. The audio and MIDI files align with an approximate 3 millisecond range. There are three versions of MAESTRO: V1.0.0, V2.0.0, and V3.0.0, which are of the sizes 87 GB, 103 GB, and 101 GB, respectively.
Slakh2100 [123]	Multi-instrument	The Synthesized Lakh (Slakh) Dataset comprises multi-track audio and synchronized MIDI data intended for music source separation and the automated transcription of multiple instruments. Utilizing high-quality sample-based virtual instruments, individual MIDI tracks are generated from the Lakh MIDI Dataset v0.1. These MIDI tracks are then combined to create musical mixtures, resulting in 2100 tracks and corresponding aligned MIDI files in this release, known as Slakh2100. These tracks are synthesized from 187 patches categorized into 34 classes.
MusicNet [124]	Multi-instrument	MusicNet is a publicly available dataset for music feature representation learning, containing 330 classical music recordings with aligned labels. These labels encompass 1,299,329 individual segments within 34 h of recordings and are based on 513 distinct instrument/note combinations. The dataset is skewed towards Beethoven's compositions and Solo Piano due to their popularity, but researchers can augment it to enhance the coverage of less-represented instruments like Flute and Oboe.
URMP [125]	Multi-instrument	URMP, the University of Rochester Multimodal Music Performance dataset, offers recordings of 44 chamber music performances, serving as a valuable resource for understanding and transcribing interactions between multiple instruments in ensemble contexts.

9. Future Directions

In the rapidly evolving landscape of music transcription, the future directions of this review paper hold promising avenues for advancing the field. Several key areas emerge as potential focal points for further investigation and development.

First, the integration of advanced machine learning techniques, such as deep learning and neural networks, offers the potential to enhance the accuracy and robustness of music transcription algorithms. The power of the transformer model is currently being

used in many different fields of artificial intelligence including automatic music transcription. Inspired by the successful sequence-to-sequence transfer learning in natural language processing, one of the recent works demonstrates the effectiveness of a general-purpose transformer model in transcribing various combinations of instruments across multiple datasets [31]. Additionally, another study takes the transformer model into account for the purpose of piano transcription. Through studies on the MAESTRO dataset and cross-dataset evaluations on the MAPS dataset, the work validates the effectiveness of the transformer model in velocity detection tasks, resulting in performance gains in both frame-level and note-level measures [126]. Exploring novel architectures, data augmentation strategies, and transfer learning approaches can contribute to more effective models capable of handling complex musical structures, including tempo changes, expressive techniques, and harmonies. Second, the seeking of real-time and interactive transcription systems holds great significance for practical applications. The design and optimization of algorithms that enable instantaneous music transcription for live performances, interactive music software, and real-time feedback systems are areas that need to be explored. Addressing challenges, developing efficient algorithms, and considering user-centered design principles will be essential in shaping the future of real-time music transcription. Moreover, the cross-disciplinary collaboration between musicologists, signal processing experts, and human–computer interaction researchers holds promise for advancing transcription methodologies. Integrating musicological insights, cognitive models of human perception, and interactive user interfaces can result in more accurate and context-aware transcription systems that align with human musical understanding and creativity. Furthermore, as music is diverse across genres, cultures, and styles, the adaptation of transcription techniques to accommodate this diversity becomes a vital endeavor. Exploring cross-cultural music transcription, genre-specific challenges, and the incorporation of non-Western musical elements can lead to more inclusive and versatile transcription tools.

10. Conclusions

This comprehensive review paper has traversed through the complex domain of music transcription by highlighting its diverse significance and challenges across various dimensions. We have explored the fundamental methodologies in four levels of transcription: frame-level transcription, note-level transcription, stream-level transcription, and notation-level transcription, emphasizing their distinct roles in musical compositions. We have explored insights and advancements in transcription techniques including general melody extraction, vocal melody extraction, multi-pitch estimation, solo vocals, ensemble vocals, as well as single-instrument and multi-instrument transcription. Similarly, we listed singing techniques, instrumentation, pitch ambiguity, tempo changes, and subjectivity as the major challenges in music transcription. Furthermore, we have explored several software tools and python libraries that can be used for different levels of transcription. Additionally, we have described the currently available music transcription datasets and their proposed uses.

This review paper serves as a guide for researchers, practitioners, and enthusiasts for the evolving landscape of music transcription. The future calls for utilizing the potential of machine learning, real-time applications, and interdisciplinary collaborations as technology innovation continues to amplify our analytical capacities. The smooth integration of computational capability and musical sensitivity holds the potential to close the gap between the auditory and the symbolic, bringing up new opportunities for music synthesis, analysis, and innovation.

Author Contributions: The first author, B.B., contributed to the whole project, which includes conceptualization, methodology, validation, analysis, and writing the original draft. The corresponding author, J.L., contributed to funding acquisition, conceptualization, and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Research Foundation of Korea (NRF) under the Development of AI for Analysis and Synthesis of Korean Pansori Project (NRF-2021R1A2C2006895).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors without undue reservation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kilmer, A.D. The discovery of an ancient Mesopotamian theory of music. *Proc. Am. Philos. Soc. USA* **1971**, *115*, 131–149.
2. West, M.L. *Ancient Greek Music*; Clarendon Press: Oxford, UK, 1992.
3. Barker, A. (Ed.) *Greek Musical Writings: Volume 2, Harmonic and Acoustic Theory*; Cambridge University Press: Cambridge, UK, 1984.
4. Hagel, S. *Ancient Greek Music: A New Technical History*; Cambridge University Press: Cambridge, UK, 2009.
5. Randel, D.M. (Ed.) *The Harvard Dictionary of Music*; Harvard University Press: Cambridge, MA, USA, 2003.
6. Leech-Wilkinson, D. *The Changing Sound of Music: Approaches to Studying Recorded Musical Performances*; Centre for the History and Analysis of Recorded Music: London, UK, 2009.
7. Gould, E. *Behind Bars: The Definitive Guide to Music Notation*; Faber Music Ltd.: London, UK, 2016.
8. Nettl, B. *The Study of Ethnomusicology: Thirty-Three Discussions*; University of Illinois Press: Champaign, IL, USA, 2015.
9. Titon, J.T. *Worlds of Music: An Introduction to the Music of the World's Peoples*; Cengage Learning: Boston, MA, USA, 2016.
10. Waring, W.; Rousseau, J.J. *A Dictionary of Music. Translated from the French of Mons. J.J. Rousseau*; J. French: London, UK, 1775.
11. Roads, C. *Composing Electronic Music: A New Aesthetic*; Oxford University Press: Oxford, UK, 2015.
12. Sigtia, S.; Benetos, E.; Dixon, S. An End-to-End Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2016**, *24*, 927–939. [CrossRef]
13. Klapuri, A.; Davy, M. (Eds.) *Signal Processing Methods for Music Transcription*; Springer-Verlag: New York, NY, USA, 2006.
14. Klapuri, A. Introduction to music transcription. In *Signal Processing Methods for Music Transcription*; Springer: Boston, MA, USA, 2006; pp. 3–20.
15. Cemgil, A.; Kappen, H.; Barber, D. A generative model for music transcription. *IEEE Trans. Audio Speech Lang. Process.* **2006**, *14*, 679–694. [CrossRef]
16. Ryynanen, M.P.; Klapuri, A. Polyphonic music transcription using note event modeling. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 16–19 October 2005; pp. 319–322.
17. Dessein, A.; Cont, A.; Lemaitre, G. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In Proceedings of the ISMIR—11th International Society for Music Information Retrieval Conference, Utrecht, The Netherlands, 9–13 August 2010; pp. 489–494.
18. Mauch, M.; Cannam, C.; Bittner, R.; Fazekas, G.; Salamon, J.; Dai, J.; Bello, J.; Dixon, S. *Computer-Aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency*; University of London: London, UK, 2015.
19. Mauch, M.; Dixon, S. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), Florence, Italy, 4–9 May 2014; pp. 659–663.
20. De Cheveigné, A.; Kawahara, H. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* **2002**, *111*, 1917–1930. [CrossRef] [PubMed]
21. Dong, H.W.; Hsiao, W.Y.; Yang, Y.H. Pypianoroll: Open source Python package for handling multitrack pianoroll. In Proceedings of the ISMIR—19th International Society for Music Information Retrieval Conference, Paris, France, 23–27 September 2018.
22. Hsu, J.Y.; Su, L. VOCANO: A Note Transcription Framework for Singing Voice in Polyphonic Music. In Proceedings of the ISMIR—22nd International Society for Music Information Retrieval Conference, Online, 7–12 November 2021; pp. 293–300.
23. Emiya, V.; Badeau, R.; David, B. Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *18*, 1643–1654. [CrossRef]
24. Su, L.; Yang, Y.-H. Combining Spectral and Temporal Representations for Multipitch Estimation of Polyphonic Music. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 1600–1612. [CrossRef]
25. Kelz, R.; Dorfer, M.; Korzeniowski, F.; Böck, S.; Arzt, A.; Widmer, G. On the potential of simple framewise approaches to piano transcription. *arXiv* **2016**, arXiv:1612.05153.
26. Berg-Kirkpatrick, T.; Andreas, J.; Klein, D. Unsupervised transcription of piano music. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1538–1546.
27. Ewert, S.; Plumbley, M.D.; Sandler, M. A dynamic programming variant of non-negative matrix deconvolution for the transcription of struck string instruments. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015), South Brisbane, Australia, 19–24 April 2015; pp. 569–573.

28. Kameoka, H.; Nishimoto, T.; Sagayama, S. A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 982–994. [CrossRef]
29. Nam, J.; Ngiam, J.; Lee, H.; Slaney, M. A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations. In Proceedings of the ISMIR—12th International Society for Music Information Retrieval Conference, Miami, FL, USA, 24–28 October 2011; pp. 175–180.
30. Boulanger-Lewandowski, N.; Bengio, Y.; Vincent, P. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv* **2012**, arXiv:1206.6392.
31. Gardner, J.; Simon, I.; Manilow, E.; Hawthorne, C.; Engel, J. MT3: Multi-task multitrack music transcription. *arXiv* **2022**, arXiv:2111.03017.
32. Duan, Z.; Han, J.; Pardo, B. Multi-pitch Streaming of Harmonic Sound Mixtures. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2013**, *22*, 138–150. [CrossRef]
33. Benetos, E.; Dixon, S. Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. *J. Acoust. Soc. Am.* **2013**, *133*, 1727–1741. [CrossRef] [PubMed]
34. Arora, V.; Behera, L. Multiple F0 Estimation and Source Clustering of Polyphonic Music Audio Using PLCA and HMRFs. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 278–287. [CrossRef]
35. Music Transcription with Convolutional Neural Networks. Available online: https://www.lunaverus.com/cnn (accessed on 5 October 2023).
36. Hsieh, T.H.; Su, L.; Yang, Y.H. A streamlined encoder/decoder architecture for melody extraction. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019), Brighton, UK, 12–17 May 2019; pp. 156–160.
37. Eggink, J.; Brown, G.J. Extracting Melody Lines from Complex Audio. In Proceedings of the ISMIR—5th International Conference on Music Information Retrieval, Barcelona, Spain, 10–14 October 2004.
38. Bittner, R.M.; McFee, B.; Salamon, J.; Li, P.; Bello, J.P. Deep Salience Representations for F0 Estimation in Polyphonic Music. In Proceedings of the ISMIR—International Society for Music Information Retrieval Conference, Suzhou, China, 23–27 October 2017; pp. 63–70.
39. Salamon, J.; Gomez, E.; Ellis, D.P.W.; Richard, G. Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges. *IEEE Signal Process. Mag.* **2014**, *31*, 118–134. [CrossRef]
40. Salamon, J.; Urbano, J. Current Challenges in the Evaluation of Predominant Melody Extraction Algorithms. In Proceedings of the ISMIR–13th International Society for Music Information Retrieval Conference, Porto, Portugal, 8–12 October 2012; Volume 12, pp. 289–294.
41. Gao, Y.; Zhang, X.; Li, W. Vocal Melody Extraction via HRNet-Based Singing Voice Separation and Encoder-Decoder-Based F0 Estimation. *Electronics* **2021**, *10*, 298. [CrossRef]
42. Lu, W.T.; Su, L. Vocal Melody Extraction with Semantic Segmentation and Audio-symbolic Domain Transfer Learning. In Proceedings of the ISMIR—19th International Society for Music Information Retrieval Conference, Paris, France, 23–27 September 2018; pp. 521–528.
43. Su, L. Vocal melody extraction using patch-based CNN. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada, 15–20 April 2018; pp. 371–375.
44. Kum, S.; Lin, J.H.; Su, L.; Nam, J. Semi-supervised learning using teacher-student models for vocal melody extraction. *arXiv* **2020**, arXiv:2008.06358.
45. Kum, S.; Nam, J. Joint Detection and Classification of Singing Voice Melody Using Convolutional Recurrent Neural Networks. *Appl. Sci.* **2019**, *9*, 1324. [CrossRef]
46. Yeh, C.; Roebel, A. Multiple-F0 Estimation for Mirex 2009. In Proceedings of the ISMIR—10th International Society for Music Information Retrieval Conference, Utrecht, The Netherlands, 26–30 October 2009; p. 1.
47. Klapuri, A.P. A perceptually motivated multiple-f0 estimation method. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 16–19 October 2005; pp. 291–294.
48. Cañadas Quesada, F.J.; Ruiz Reyes, N.; Vera Candeas, P.; Carabias, J.J.; Maldonado, S. A multiple-F0 estimation approach based on Gaussian spectral modelling for polyphonic music transcription. *J. New Music. Res.* **2010**, *39*, 93–107. [CrossRef]
49. Chang, W.C.; Su, A.W.; Yeh, C.; Roebel, A.; Rodet, X. Multiple-F0 tracking based on a high-order HMM model. In *Digital Audio Effects (DAFx-08)*; HAL: Bangalore, India, 2008.
50. Christensen, M.; Jakobsson, A. *Multi-Pitch Estimation*; Synthesis Lectures on Speech and Audio Processing Series; Morgan and Claypool: San Rafael, CA, USA, 2009.
51. Cuesta, H.; McFee, B.; Gómez, E. Multiple f0 estimation in vocal ensembles using convolutional neural networks. In Proceedings of the ISMIR—21th International Society for Music Information Retrieval Conference, virtual conference, 11–16 October 2020; pp. 302–309.
52. Yang, L.; Maezawa, A.; Smith, J.B.; Chew, E. Probabilistic transcription of sung melody using a pitch dynamic model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orlean, LA, USA, 5–9 March 2017; pp. 301–305.
53. Meseguer-Brocal, G.; Bittner, R.; Durand, S.; Brost, B. Data cleansing with contrastive learning for vocal note event annotations. *arXiv* **2020**, arXiv:2008.02069.

54. Molina, E.; Tardon, L.J.; Barbancho, A.M.; Barbancho, I. SiPTH: Singing Transcription Based on Hysteresis Defined on the Pitch-Time Curve. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *23*, 252–263. [CrossRef]
55. Bittner, R.M.; Pasalo, K.; Bosch, J.J.; Meseguer-Brocal, G.; Rubinstein, D. vocadito: A dataset of solo vocals with f_0, note, and lyric annotations. *arXiv* **2021**, arXiv:2110.05580.
56. Ternström, S. Choir acoustics: An overview of scientific research published to date. *Int. J. Res. Choral Sing.* **2003**, *1*, 3–12.
57. Dai, J.; Dixon, S. Analysis of Interactive Intonation in Unaccompanied SATB Ensembles. In Proceedings of the ISMIR—18th International Society for Music Information Retrieval Conference, Suzhou, China, 23–27 October 2017; pp. 599–605.
58. Schramm, R.; Benetos, E. Automatic Transcription of a Cappella Recordings from Multiple Singers. In Proceedings of the Audio Engineering Society Conference, Arlington, VA, USA, 15–17 June 2017; Available online: https://www.proceedings.com/audio-engineering-society-aes/ (accessed on 5 October 2023).
59. Ternström, S. Perceptual evaluations of voice scatter in unison choir sounds. *J. Voice* **1993**, *7*, 129–135. [CrossRef] [PubMed]
60. McLeod, A.; Schramm, R.; Steedman, M.; Benetos, E. Automatic Transcription of Polyphonic Vocal Music. *Appl. Sci.* **2017**, *7*, 1285. [CrossRef]
61. Schramm, R. Automatic transcription of polyphonic vocal music. In *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 715–735.
62. Hawthorne, C.; Elsen, E.; Song, J.; Roberts, A.; Simon, I.; Raffel, C.; Engel, J.; Oore, S.; Eck, D. Onsets and frames: Dual-objective piano transcription. *arXiv* **2017**, arXiv:1710.11153.
63. Kong, Q.; Li, B.; Song, X.; Wan, Y.; Wang, Y. High-Resolution Piano Transcription with Pedals by Regressing Onset and Offset Times. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3707–3717. [CrossRef]
64. Cheng, T.; Mauch, M.; Benetos, E.; Dixon, S. An attack/decay model for piano transcription. In Proceedings of the ISMIR–17th International Society for Music Information Retrieval Conference, New York, NY, USA, 7–11 August 2016; pp. 584–590.
65. Koepke, A.S.; Wiles, O.; Moses, Y.; Zisserman, A. Sight to sound: An end-to-end approach for visual piano transcription. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 4–8 May 2020; pp. 1838–1842.
66. Hawthorne, C.; Stasyuk, A.; Roberts, A.; Simon, I.; Huang, C.Z.; Dieleman, S.; Elsen, E.; Engel, J.; Eck, D. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In ICLR. *arXiv* **2019**, arXiv:1810.12247.
67. Fiss, X.; Kwasinski, A. Automatic real-time electric guitar audio transcription. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), Prague, Czech Republic, 22–27 May 2011; pp. 373–376.
68. Kim, S.; Hayashi, T.; Toda, T. Note-level automatic guitar transcription using attention mechanism. In Proceedings of the 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022; pp. 229–233.
69. Barbancho, I.; de la Bandera, C.; Barbancho, A.M.; Tardon, L.J. Transcription and expressiveness detection system for violin music. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 189–192.
70. Maezawa, A.; Itoyama, K.; Komatani, K.; Ogata, T.; Okuno, H.G. Automated Violin Fingering Transcription Through Analysis of an Audio Recording. *Comput. Music. J.* **2012**, *36*, 57–72. [CrossRef]
71. Al-Ghawanmeh, F.M.; Jafar, I.F.; Al-taee, M.A.; Al-ghawanmeh, M.T.; Muhsin, Z.J. Development of Improved automatic music transcription system for the Arabian flute (nay). In Proceedings of the 8th International Multi-Conference on Systems, Signals & Devices, Sousse, Tunisia, 22–25 March 2011; pp. 1–6.
72. Al-Taee, M.A.; Al-Rawi, M.S.; Al-Ghawanmeh, F.M. Time-frequency analysis of the Arabian flute (Nay) tone applied to automatic music transcription. In Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications, Doha, Qatar, 31 March–4 April 2008; pp. 891–894.
73. Cheuk, K.W.; Choi, K.; Kong, Q.; Li, B.; Won, M.; Hung, A.; Wang, J.C.; Herremans, D. Jointist: Joint learning for multi-instrument transcription and its applications. *arXiv* **2022**, arXiv:2206.10805.
74. Wu, Y.-T.; Chen, B.; Su, L. Multi-Instrument Automatic Music Transcription with Self-Attention-Based Instance Segmentation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2020**, *28*, 2796–2809. [CrossRef]
75. Tanaka, K.; Nakatsuka, T.; Nishikimi, R.; Yoshii, K.; Morishima, S. Multi-Instrument Music Transcription Based on Deep Spherical Clustering of Spectrograms and Pitchgrams. In Proceedings of the ISMIR—21th International Society for Music Information Retrieval Conference, virtual conference, 11–16 October 2020; pp. 327–334.
76. Wu, Y.T.; Chen, B.; Su, L. Polyphonic music transcription with semantic segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2018), Brighton, UK, 12–17 May 2019; pp. 166–170.
77. Vogl, R.; Widmer, G.; Knees, P. Towards multi-instrument drum transcription. *arXiv* **2018**, arXiv:1806.06676.
78. Cheuk, K.W.; Choi, K.; Kong, Q.; Li, B.; Won, M.; Wang, J.C.; Herremans, Y.N. Jointist: Simultaneous Improvement of Multi-instrument Transcription and Music Source Separation via Joint Training. *arXiv* **2023**, arXiv:2302.00286.
79. Manilow, E.; Seetharaman, P.; Pardo, B. Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 4–8 May 2020; pp. 771–775.
80. Yamamoto, Y.; Nam, J.; Terasawa, H. PrimaDNN': A Characteristics-aware DNN Customization for Singing Technique Detection. European signal processing conference (EUSIPCO). *arXiv* **2023**, arXiv:2306.14191.

81. Yamamoto, Y.; Nam, J.; Terasawa, H. Analysis and Detection of Singing Techniques in Repertoires of J-POP Solo Singers. In Proceedings of the ISMIR—23th International Society for Music Information Retrieval Conference, Bengaluru, India, 4–8 December 2022.
82. Heidemann, K.A. System for Describing Vocal Timbre in Popular Song. *J. Soc. Music Theory* **2016**, *22*, 1–17. [CrossRef]
83. Emmons, S.; Chase, C. *Prescriptions for Choral Excellence*; Oxford University Press: Oxford, UK, 2006.
84. Doscher, B.M. *The Functional Unity of the Singing Voice*; Scarecrow Press: Lanham, MD, USA, 1993.
85. Kanno, M. Thoughts on how to play in tune: Pitch and intonation. *Contemp. Music. Rev.* **2003**, *22*, 35–52. [CrossRef]
86. Huang, H.; Huang, R. She Sang as She Spoke: Billie Holiday and Aspects of Speech Intonation and Diction. *Jazz Perspect.* **2013**, *7*, 287–302. [CrossRef]
87. McAdams, S.; Goodchild, M. Musical structure: Sound and timbre. In *Routledge Companion to Music Cognition*; Routledge: London, UK, 2017; pp. 129–139.
88. Xiao, Z.; Chen, X.; Zhou, L. Polyphonic piano transcription based on graph convolutional network. *Signal Process.* **2023**, *212*, 109134. [CrossRef]
89. Yang, D.; Tsai, T.J. Piano sheet music identification using dynamic n-gram fingerprinting. *Trans. Int. Soc. Music. Inf. Retr.* **2021**, *1*, 4. [CrossRef]
90. Hasanain, A.; Syed, M.; Kepuska, V.; Silaghi, M. Multi-Dimensional Spectral Process for Cepstral Feature Engineering & Formant Coding. *J. Electr. Electron. Eng.* **2022**, *1*, 1–20.
91. Korzeniowski, F.; Widmer, G. Feature learning for chord recognition: The deep chroma extractor. *arXiv* **2016**, arXiv:1612.05065.
92. Repp, B.H.; Luke Windsor, W.; Desain, P. Effects of tempo on the timing of simple musical rhythms. *Music. Percept.* **2002**, *19*, 565–593. [CrossRef]
93. Quinn, S.; Watt, R. The perception of tempo in music. *Perception* **2006**, *35*, 267–280. [CrossRef]
94. Koops, H.V.; de Haas, W.B.; Burgoyne, J.A.; Bransen, J.; Kent-Muller, A.; Volk, A. Annotator subjectivity in harmony annotations of popular music. *J. New Music. Res.* **2019**, *48*, 232–252. [CrossRef]
95. Kurth, F.; Müller, M.; Fremerey, C.; Chang, Y.H.; Clausen, M. Automated Synchronization of Scanned Sheet Music with Audio Recordings. In Proceedings of the ISMIR—8th International Society for Music Information Retrieval Conference, Vienna, Austria, 23–27 September 2007; pp. 261–266.
96. Bereket, M.; Shi, K. *An AI Approach to AutZomatic Natural Music Transcription*; Stanford University: Stanford, CA, USA, 2017.
97. Doremir Music Research AB, ScoreCloud. Available online: https://www.audacityteam.org/ (accessed on 5 October 2023).
98. Audacity Software. 1999. Available online: http://thurs3.pbworks.com/f/audacity.pdf (accessed on 5 October 2023).
99. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference 2015, Austin, TX, USA, 6–12 July 2015; Volume 8, pp. 18–25.
100. Cuthbert, M.S.; Ariza, C. music21: A toolkit for computer-aided musicology and symbolic music data. In Proceedings of the ISMIR—11th International Society for Music Information Retrieval Conference, Utrecht, The Netherlands, 9–13 August 2010.
101. Böck, S.; Korzeniowski, F.; Schlüter, J.; Krebs, F.; Widmer, G. Madmom: A new python audio and music signal processing library. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 1 October 2016; pp. 1174–1178.
102. Bogdanov, D.; Wack, N.; Gómez Gutiérrez, E.; Gulati, S.; Boyer, H.; Mayor, O.; Roma Trepat, G.; Salamon, J.; Zapata González, J.R.; Serra, X. Essentia: An audio analysis library for music information retrieval. In Proceedings of the ISMIR–14th International Society for Music Information Retrieval Conference, Curitiba, Brazil, 4–8 November 2013; pp. 493–498.
103. Brossier, P.M. The aubio library at mirex 2006. In *Synthesis*; Queen Mary University of London: London, UK, 2006.
104. Dittmar, C.; Cano, E.; Abeber, J.; Grollmisch, S. Music information retrieval meets music education. In *Dagstuhl Follow-Ups*; Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik: Wadern, Germany, 2012; Volume 3.
105. Wang, Y.; Zhang, B. Application-specific music transcription for tutoring. *IEEE MultiMedia* **2008**, *15*, 70–74. [CrossRef]
106. Sturm, B.L.; Ben-Tal, O.; Monaghan, U.; Collins, N.; Herremans, D.; Chew, E.; Hadjeres, G.; Deruty, E.; Pachet, F. Machine learning research that matters for music creation: A case study. *J. New Music. Res.* **2018**, *48*, 36–55. [CrossRef]
107. Goto, M.; Dannenberg, R.B. Music Interfaces Based on Automatic Music Signal Analysis: New Ways to Create and Listen to Music. *IEEE Signal Process. Mag.* **2018**, *36*, 74–81. [CrossRef]
108. Bocko, G.; Bocko, M.F.; Headlam, D.; Lundberg, J.; Ren, G. Automatic music production system employing probabilistic expert systems. In *Audio Engineering Society Convention*; Audio Engineering Society: Dearborn, MI, USA, 2010.
109. Benetos, E.; Dixon, S.; Giannoulis, D.; Kirchhoff, H.; Klapuri, A. Automatic music transcription: Challenges and future directions. *J. Intell. Inf. Syst.* **2013**, *41*, 407–434. [CrossRef]
110. Weyde, T.; Cottrell, S.; Dykes, J.; Benetos, E.; Wolff, D.; Tidhar, D.; Kachkaev, A.; Plumbley, M.; Dixon, S.; Barthet, M.; et al. Big data for musicology. In Proceedings of the 1st International Workshop on Digital Libraries for Musicology, London, UK, 12 September 2014; pp. 1–3.
111. Zhang, L.; Li, R.; Wang, S.; Deng, L.; Liu, J.; Ren, Y.; He, J.; Huang, R.; Zhu, J.; Chen, X.; et al. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 6914–6926.
112. Ryynanen, M.; Virtanen, T.; Paulus, J.; Klapuri, A. Accompaniment separation and karaoke application based on automatic melody transcription. In Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, 23–26 June 2008; pp. 1417–1420.

113. Mostafa, N.; Wan, Y.; Amitabh, U.; Fung, P. A machine learning based music retrieval and recommendation system. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 1970–1977.
114. Tachibana, H.; Mizuno, Y.; Ono, N.; Sagayama, S. A real-time audio-to-audio karaoke generation system for monaural recordings based on singing voice suppression and key conversion techniques. *J. Inf. Process.* **2016**, *24*, 470–482. [CrossRef]
115. Sripradha, R.; Surya, P.; Supraja, P.; Manitha, P.V. Karaoke Machine Execution Using Artificial Neural Network. In *Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAAIML*; Springer: Singapore, 2020; pp. 693–704.
116. Wu, Y.-T.; Luo, Y.-J.; Chen, T.-P.; Wei, I.-C.; Hsu, J.-Y.; Chuang, Y.-C.; Su, L. Omnizart: A General Toolbox for Automatic Music Transcription. *J. Open Source Softw.* **2021**, *6*, 3391. [CrossRef]
117. Hsu, C.-L.; Jang, J.-S.R. On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *18*, 310–319.
118. Bittner, R.M.; Salamon, J.; Tierney, M.; Mauch, M.; Cannam, C.; Bello, J.P. Medleydb: A multitrack dataset for annotation-intensive mir research. In Proceedings of the ISMIR–15th International Society for Music Information Retrieval Conference, Taipei, Taiwan, 27–31 October 2014; pp. 155–160.
119. Duan, Z.; Pardo, B.; Zhang, C. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *18*, 2121–2133. [CrossRef]
120. Gómez, E.; Bonada, J. Towards Computer-Assisted Flamenco Transcription: An Experimental Comparison of Automatic Transcription Algorithms as Applied to A Cappella Singing. *Comput. Music. J.* **2013**, *37*, 73–90. [CrossRef]
121. Condit-Schultz, N.; Ju, Y.; Fujinaga, I. A Flexible Approach to Automated Harmonic Analysis: Multiple Annotations of Chorales by Bach and Prætorius. In Proceedings of the ISMIR—19th International Society for Music Information Retrieval Conference, Paris, France, 23–27 September 2018; pp. 66–73.
122. Gong, X.; Xu, W.; Liu, J.; Cheng, W. Analysis and correction of maps dataset. In Proceedings of the 22th International Conference on Digital Audio Effects (DAFx-19), Birmingham, UK, 2–6 September 2019.
123. Manilow, E.; Wichern, G.; Seetharaman, P.; Le Roux, J. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In Proceedings of the Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2019), New Paltz, NY, USA, 20–23 October 2019; pp. 45–49.
124. Thickstun, J.; Harchaoui, Z.; Kakade, S. Learning features of music from scratch. *arXiv* **2017**, arXiv:1611.09827.
125. Li, B.; Liu, X.; Dinesh, K.; Duan, Z.; Sharma, G. Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. *IEEE Trans. Multimed.* **2018**, *21*, 522–535. [CrossRef]
126. Ou, L.; Guo, Z.; Benetos, E.; Han, J.; Wang, Y. Exploring transformer's potential on automatic piano transcription. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022), Singapore, 22–27 May 2022; pp. 776–780.