A music cognition-guided framework for multi-pitch estimation.

LI, X., YAN, Y., SORAGHAN, J., WANG, Z. and REN, J.

2023

Cognitive Computation (2023) 15:23–35
https://doi.org/10.1007/s12559-022-10031-5

A Music Cognition–Guided Framework for Multi-pitch Estimation

Xiaoquan Li[1] · Yijun Yan[2] · John Soraghan[1] · Zheng Wang[3] · Jinchang Ren[2]

Received: 3 September 2021 / Accepted: 26 May 2022 / Published online: 14 June 2022

Abstract
As one of the most important subtasks of automatic music transcription (AMT), multi-pitch estimation (MPE) has been studied extensively for predicting the fundamental frequencies in the frames of audio recordings during the past decade. However, how to use music perception and cognition for MPE has not yet been thoroughly investigated. Motivated by this, this demonstrates how to effectively detect the fundamental frequency and the harmonic structure of polyphonic music using a cognitive framework. Inspired by cognitive neuroscience, an integration of the constant Q transform and a state-of-the-art matrix factorization method called shift-invariant probabilistic latent component analysis (SI-PLCA) are proposed to resolve the polyphonic short-time magnitude log-spectra for multiple pitch estimation and source-specific feature extraction. The cognitions of rhythm, harmonic periodicity and instrument timbre are used to guide the analysis of characterizing contiguous notes and the relationship between fundamental frequency and harmonic frequencies for detecting the pitches from the outcomes of SI-PLCA. In the experiment, we compare the performance of proposed MPE system to a number of existing state-of-the-art approaches (seven weak learning methods and four deep learning methods) on three widely used datasets (i.e. MAPS, BACH10 and TRIOS) in terms of F-measure (F_1) values. The experimental results show that the proposed MPE method provides the best overall performance against other existing methods.

Keywords Music cognition · Automatic music transcription · Multi-pitch estimation · Harmonic structure detection (HSD) · Polyphonic music detection

Introduction

Estimation and tracking of multiple fundamental frequencies is one of the major tasks in automatic music transcription (AMT) of polyphonic music analysis [1] and music information retrieval (MIR) [2], which is referred to as a subtask in the Music Information Retrieval Evaluation eXchange (MIREX).[1] Multiple fundamental frequency estimation (MFE), also namely multiple pitch estimation (MPE), is challenging in processing simultaneous notes from multiple instruments in polyphonic music [3, 4]. There is often a trade-off between the robustness and efficiency of algorithms that focuses more on complexity rather than single-pitch estimation.

According to Benetos et al. [5], the MPE approaches are categorised into three types, i.e. feature based, spectrogram-factorization based and statistical model–based methods. In feature-based methods, signal processing techniques such as the pitch salience function [6] and pitch candidate set score function [7] are used. In spectrogram-factorization methods, both the nonnegative matrix factorisation (NMF) and the probabilistic latent component analysis (PLCA) approaches have received a lot of attention in recent years [6], and numerous improved versions [8, 9] based on both methods have been published and are recognised as leading spectrogram factorization-based methods in the MPE domain. The statistical model-based methods employ the maximum a posteriori (MAP) [3] estimation, maximum likelihood (ML) or Bayesian theory [10] to detect the fundamental frequencies.. It is worth noting that these three distinct types of MPE approaches can be joined or interacted with [6] for a variety of applications.

✉ Jinchang Ren
jinchang.ren@ieee.org

[1] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK

[2] National Subsea Centre, Robert Gordon University, Aberdeen AB21 0BH, UK

[3] College of Intelligence and Computing, Tianjin University, Tianjin, China

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

In recent years, many deep learning (DL)–based supervised MPE approaches have also been developed. Cheuk et al. [11] presented a DL model for AMT by combining the U-Net and bidirectional long short-term memory (BiLSTM) neural network modules. Mukherjee et al. [12] used statistical characteristics and an extreme learning machine for musical instrument segregation, where LSTM and the recurrent neural network (RNN) [13] were combined to differentiate the musical chords for AMT. Fan et al. [14] proposed a deep neural network to extract the singing voice, followed by a dynamic unbroken pitch determination algorithm to track pitches. Sigtia et al. [15] developed a supervised approach for polyphonic piano music transcription that included a RNN and a probabilistic graphical model. Although DL approaches may provide adequate music transcriptions, they often require high-performance computers and excellent graphic processing units (GPU) to speed-up the lengthy training process [16]. Furthermore, DL algorithms may suffer from inaccurately labelled data, and the performance may be susceptible to the training samples and the learning procedures used. To this end, in this paper, we focus mainly on cognitive method, where the prior cognitive theories and assumptions from previous studies [17–19] will be used to guide the fundamental pitch detection in polyphonic music pieces.

To distinguish the pitch using harmonic analysis, two types of statistical models are often used. One is the expectation–maximization (EM)-based algorithms [20], and the other is Bayesian-based algorithms [21]. For EM-based methods, Emiya et al. [22] proposed a maximum likelihood–based method for multi-pitch estimation. Duan and Temperley [23] proposed a three-stage music transcription system and applied maximum likelihood for final note tracking. For Bayesian-based methods, Alvarado Duran [24] combined Gaussian processes and Bayesian models for multi-pitch estimation. Nishikimi et al. [25] integrate hidden Markov Model and Bayesian inference together to precisely detect the vocal pitch. Those statistical models can be also considered as shallow learning methods, as data should first be observed to gain some prior knowledge, based on which the experiments should then be conducted. After constant addition of the information of the new samples into prior distribution, the posterior inference can be delivered along with the final results. Although the shallow learning approaches have been widely investigated [26], they still have much room to improve.

Apart from the aforementioned issues, most MPE methods are designed from the viewpoint of signal processing rather than music cognition, resulting in a lack of sufficient underpinning theory and inefficient modelling. To tackle this issue, we propose a general framework in which music cognitions are used to guide the entire process of MPE. In the pre-processing, inspired by cognitive neuroscience of music [19], the Constant Q transform (CQT) [27] is employed to transfer the audio signal to time–frequency spectrogram. The pianoroll transcription is then generated using a conventional matrix factorization approach, shift-invariant probabilistic latent component analysis (SI-PLCA) [9]. In the harmonic structure detection (HSD) process, the cognitions of harmonic periodicity and instrument timbre [18] are used to guide the extraction of multiple pitches. The efficacy of the suggested methodologies has been fully validated by experiments on three publicly available datasets.

The major contributions of this paper may be highlighted as follows. First, a new HSD model that incorporates music cognition for multiple fundamental frequency extraction was proposed. Second, we proposed a new note tracking method guided by music connectivity and multi-pitch model. By combining conventional pianoroll transcription approaches and the proposed HSD model, a new music cognition–guided optimization framework is introduced for MPE. Experimental results on three datasets have demonstrated the merits of our approach, when benchmarked with 11 state-of-the-art methods.

The rest of the paper is structured as follows: "Cognition-guided multiple pitch estimation" describes pre-processing for MPE including time–frequency representation, matrix factorization and the implementation of the proposed harmonic structure detection method. "Experimental results" presents the experimental results and performance analysis. Finally, a thorough conclusion is drawn in "Conclusion".

Cognition-Guided Multiple Pitch Estimation

System Overview

The objective of this work is to detect the multiple pitches from music pieces of mixed instruments, where an MPE system is proposed, which contains three key modules, i.e., pre-processing, harmonic structure detection and note tracking. Preprocessing covers a standard procedure, in which an input music signal needs to go through time–frequency (TF) representation and matrix factorization for feature extraction. The overall diagram of the MPE framework is illustrated in Fig. 1, where the implementation details are presented as follows.

Pre-processing

According to the cognitive neuroscience of music [19, 28], before selectively stimulating the auditory cortex, different frequencies within the music need to be first filtered by human cochlea. As the frequency of human auditory perception is logarithmically distributed [27], there is a greater discrimination when hearing relatively lower frequencies.

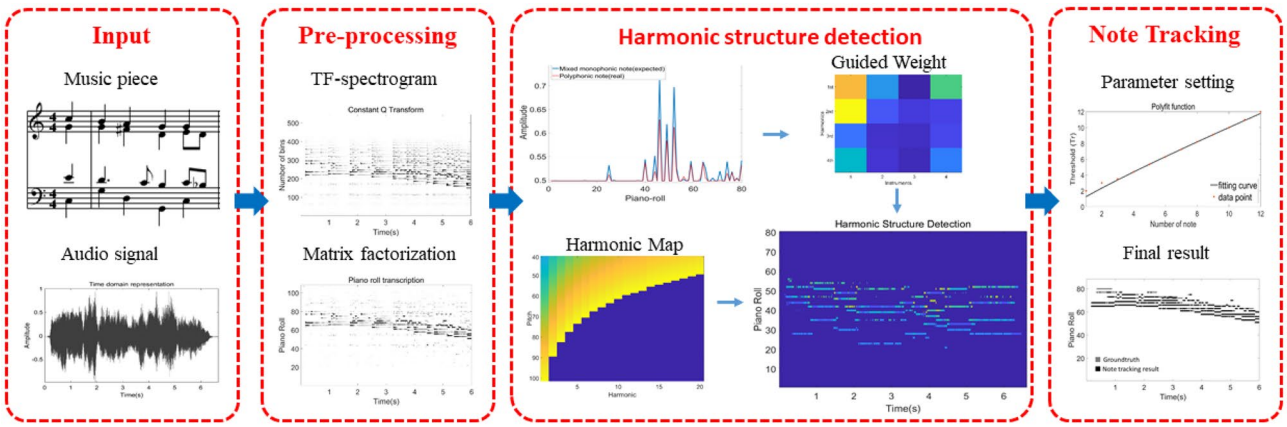

Fig. 1 The overall MPE system

The Constant Q transform (CQT) [29], based on the FFT principle, can process a logarithmic compression similar to that of human's cochlea helical structure [29]. Therefore, the CQT is employed as the TF representation module to derive the TF spectrogram, as it is efficient in lower frequencies. There are fewer frequencies required in a given range, which has testified its usefulness when the frequency distribution in several octaves is discrete. Meanwhile, an increased frequency bin correlates to a decrease in the temporal resolution rate, making it suitable for auditory applications. A spectral resolution of 60 bins per octave is used as suggested by Brown [27]. The outputs from the TF transformation are linear when using the Fast Fourier Transform (FFT) to analyse the frequency (Fig. 2a).

In the matrix factorization module, the CQT spectrogram results are used as the input, approximately modelled as a bivariate probability distribution $\boldsymbol{P}(p,t)$. The output of this module is a 2-dimensional non-binary representation of pianoroll transcription (a pitch vs. time matrix shown in Fig. 2b). In this paper, the fast shift-invariant probabilistic latent component analysis (SI-PLCA) [30] is used for automatic transcription of polyphonic music, as it is extremely useful for log-frequency spectrogram, due to the same inter-harmonic spacing for all periodic sounds [31]. Given an input signal $\boldsymbol{X}_t$, the output of CQT is a log-frequency spectrogram $\boldsymbol{V}_{z,t}$ that can be considered as a joint time–frequency distribution $\boldsymbol{P}(z,t)$ where z and t denote the frequency and time, respectively. After applying the SI-PLCA, $\boldsymbol{P}(z,t)$ can be further decomposed into several components by [30]:

$$\boldsymbol{V}_{z,t} = \boldsymbol{P}(z,t) = \boldsymbol{P}(t) \sum_{p,f,s} \boldsymbol{P}(z-f|s,p)\boldsymbol{P}_t(f|p)\boldsymbol{P}_t(s|p)\boldsymbol{P}_t(p) \quad (1)$$

where p,f,s are latent variables which denote respectively the pitch index, pitch-shifting parameter and instrument source. In Eq. (1), $\boldsymbol{P}(t)$ is the energy distribution of the spectrogram, which is known from the input signal. $\boldsymbol{P}(z-f|s,p)$ denotes the spectral templates for a given pitch

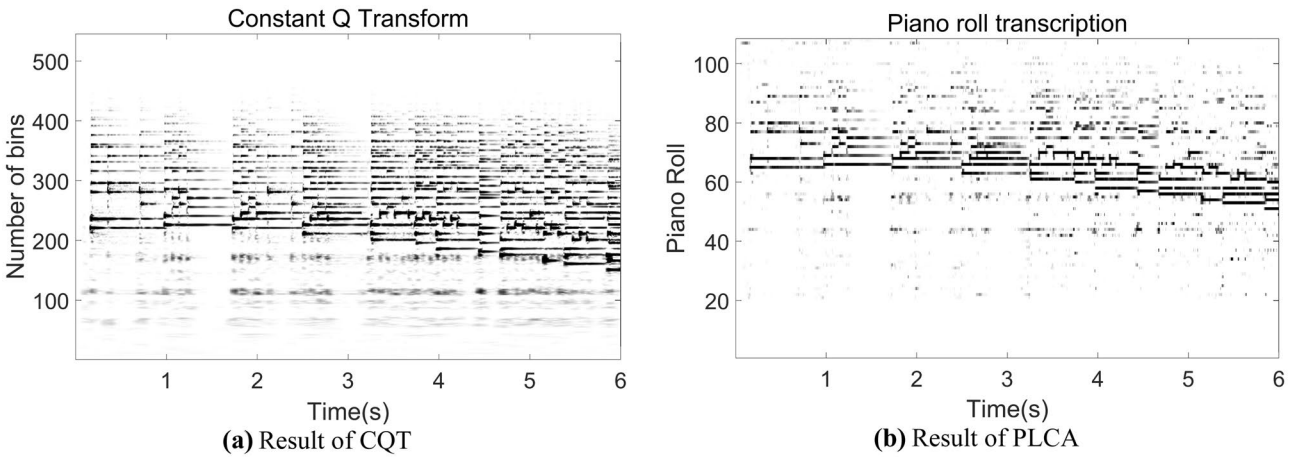

Fig. 2 **a** Result of CQT. **b** Result of PLCA. Illustration of input music signal TF representation module and pianoroll transcription module, the range from 200 to 300 bins in **a** are probably corresponding to 40–60 pitches in **b**

p and instrument source s with f pitch shifting across the log-frequency. $\boldsymbol{P}_t(f|p)$ is the log-frequency shift for each pitch on the time frame t, $\boldsymbol{P}_t(s|p)$ represents instrumentation contribution for the pitch in the time frame t, and $\boldsymbol{P}_t(p)$ is the pitch contribution which can be considered as transcription matrix on the time frame t. Since there are latent variables in this model, the expectation maximization (EM) algorithm [20] is often used to iteratively estimate the corresponding unknown variables.

In the Expectation step, the Bayes's theorem is adopted to estimate the contribution of the latent variables p, f, s for reconstruction of the model:

$$\boldsymbol{P}_t(p,f,s|z) = \frac{\boldsymbol{P}(z-f|s,p)\boldsymbol{P}_t(f|p)\boldsymbol{P}_t(s|p)\boldsymbol{P}_t(p)}{\sum_{p,f,s}\boldsymbol{P}(z-f|s,p)\boldsymbol{P}_t(f|p)\boldsymbol{P}_t(s|p)\boldsymbol{P}_t(p)} \tag{2}$$

In the Maximization step, the posterior of Eq. (2) is used to maximise the log-likelihood function in Eq. (3), which leads to the update of Eqs. (4)–(7). As suggested in [30], this step can converge after 15–20 iterations. The final result of the pianoroll transcription is derived by $\boldsymbol{P}(p,t) = \boldsymbol{P}(t)\boldsymbol{P}_t(p)$:

$$\mathcal{L} = \sum_{z,t} \boldsymbol{V}_{z,t} \log\left(\boldsymbol{P}(z,t)\right) \tag{3}$$

$$\boldsymbol{P}_t(z|s,p) = \frac{\sum_{f,t}\boldsymbol{P}_t(p,f,s|z+f)\boldsymbol{P}(z+f,t)}{\sum_{f,w,t}\boldsymbol{P}_t(p,f,s|z+f)\boldsymbol{P}(z+f,t)} \tag{4}$$

$$\boldsymbol{P}_t(f|p) = \frac{\sum_{z,s}\boldsymbol{P}_t(p,f,s|z)\boldsymbol{P}(z,t)}{\sum_{f,z,s}\boldsymbol{P}_t(p,f,s|z)\boldsymbol{P}(z,t)} \tag{5}$$

$$\boldsymbol{P}_t(s|p) = \frac{\sum_{z,f}\boldsymbol{P}_t(p,f,s|z)\boldsymbol{P}(z,t)}{\sum_{s,z,f}\boldsymbol{P}_t(p,f,s|z)\boldsymbol{P}(z,t)} \tag{6}$$

$$\boldsymbol{P}_t(p) = \frac{\sum_{z,f,s}\boldsymbol{P}_t(p,f,s|z)\boldsymbol{P}(z,t)}{\sum_{p,z,f,s}\boldsymbol{P}_t(p,f,s|z)\boldsymbol{P}(z,t)} \tag{7}$$

Harmonic Structure Detection

This section is the core of the proposed MPE system where music theories in terms of the pattern of beat length and assumption of equal energy between mixed monophonic and polyphonic music pieces are used to guide the model for the extraction of the multiple fundamental frequencies from a mixture of music sources.

For a given piece of music, the time domain representation is illustrated in the input module in Fig. 1. The results of CQT and SI-PLCA are given in Fig. 2a and b, respectively. Upon observing Fig. 2b, the fundamental pitch and its harmonics have been highlighted by the shaded black and grey strips. However, there is considerable noise and redundant information represented by small and grey dots which may be misconstrued for pitches at lower frequencies. Furthermore, the white gaps in the black and grey strips indicate frequency information that has been lost in the analysis. This suggests that the consistency of fundamental pitch is insufficient if considered frame by frame (each frame was set to 10 ms). To address these inconsistencies, the HSD method is proposed followed by a note tracking process (Fig. 1).

The proposed HSD includes two main stages. In the first stage, the pianoroll transcription $\boldsymbol{P}(p,t)$ is normalised into [0, 1] by using the following max-mean sigmoid activation function [32]:

$$\boldsymbol{PN} = \frac{1}{1+e^{-z}} \tag{8}$$

$$z = \frac{\boldsymbol{P}(p,t) - \text{mean}\left(\boldsymbol{P}(p,t)\right)}{\max\left(\boldsymbol{P}(p,t)\right) - \min\left(\boldsymbol{P}(p,t)\right)} \tag{9}$$

where $\boldsymbol{PN}$ represents the normalised $\boldsymbol{P}(p,t)$. By applying a mean filter in Eqs. (8) and (9), the spectrogram can be smoothed. For extreme values which are too large or too small than expected, they can also be rationalised. For any $\boldsymbol{PN}$, the value of $\boldsymbol{PN}_t$ at time t can be expressed by Eq. (10).

$$\boldsymbol{PN}_t = (\boldsymbol{PN}_{t-1} + \boldsymbol{PN}_t + \boldsymbol{PN}_{t+1})/3 \tag{10}$$

$$\boldsymbol{PF}_t = \boldsymbol{PN}_t \times s;\ s = \begin{cases} 1, & \textit{if } \boldsymbol{PN} > TH_1 \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

Inspired by the music theory that most high-order harmonic components are in the high-frequency range with low amplitude [17], a two-step hard constrain is used to remove most of the high-frequency components, noise and redundancy. First, a fixed threshold TH_1 is applied in Eq. (11) to remove small values. Based on the characteristic of sigmoid function (Eq. (8)), TH_1 is set to 0.5. Finally, the filtered result $\boldsymbol{PF}$ of the whole frames is obtained and shown in Fig. 3a.

In the second step, the statistics of the beat length is used to guide the removal of noise and redundant information. According to the cognition of music perception, most notes in musical rhythms have a large number of crotchets and quavers, but fewer numbers of semiquavers and demisemiquavers [33]. The rate of occurrence of different notes in the BACH10 database was observed and measured according to the ground truth. A plot was generated of time vs. rate of occurrence in Fig. 4, with the labelled fractions (i.e. $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$) denoting minim, crotchet, quaver, semiquaver and demisemiquaver, respectively. Figure 4 illustrates that the rate of occurrence of

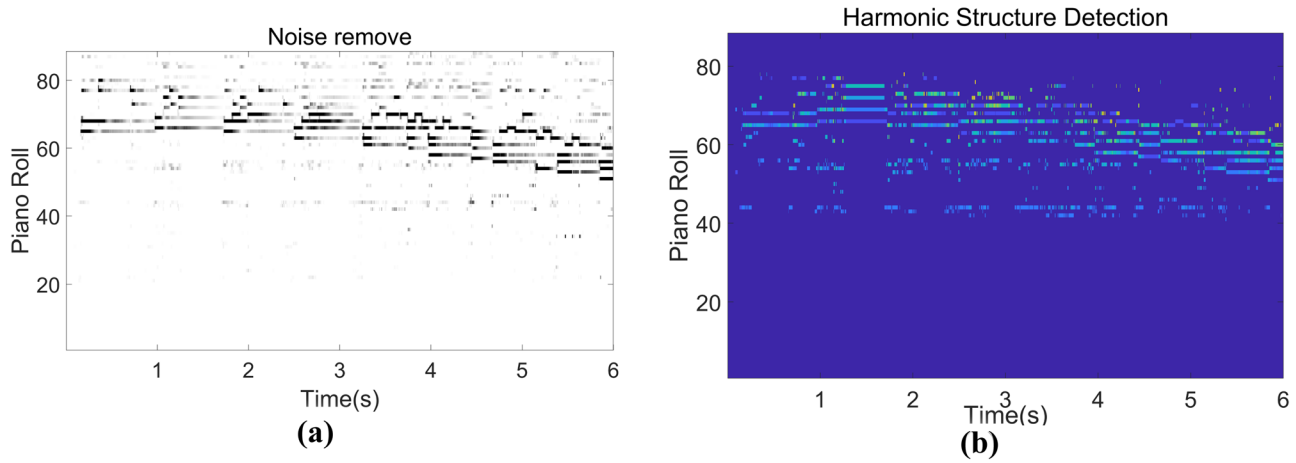

Fig. 3 Results from the first step (**a**) and second step (**b**) in HSD module

crotchets and quavers is larger than that of the demisemiquavers, semiquavers and minims. Especially, the number of demisemiquavers and semiquavers is extremely low. Furthermore, if the length of a semibreve is defined as τ, the length of a demisemiquaver is $\tau/32$. Any notes shorter than a demisemiquaver will be removed in ***PF*** before any further processing in the second stage.

In Fig. 4, a peak value is identified at the initial time steps of the simulation, and this may be due to two reasons. Firstly, manually played music may contain some timing errors, for example, holding a note for its precise duration for every note in the piece may be impossible. Secondly, ornaments such as vibrato and glissando may be mistakenly performed despite not being present on the music score. The length of such vibrato and glissando is equal to a demisemiquaver or lower [34]. To extract more of the main body of multiple pitches, factors such as human playing habits or ornaments are ignored in the proposed work. Relevant results given in "Experimental results" demonstrate that the multiple pitches are highlighted whilst removing most of the unwanted noise.

After filtering the amplitudes from PLCA, the HSD framework was proposed to detect the fundamental pitch in the second stage. The flowchart in Fig. 5 outlines the process of HSD, and Table 1 lists the description of each parameter. As described in the flowchart in Fig. 5, the output from previous steps will be analysed in two domains, i.e. pitch domain ***PD*** and energy domain ***ED***. In this context, each frame of ***PF*** is split into two vectors, ***PD***(n) and ***ED***(n). ***PD***$(n)\epsilon\mathbb{R}^{N*1}$ is non-zero notes index in each frame, ***ED***$(n)\epsilon\mathbb{R}^{N*1}$ is the amplitude of ***PD***(n), and N is the number of non-zero notes. As seen, the process is only applied once on the non-zero notes rather than the whole frame, because there is no need to analyse those zero-value notes for efficiency.

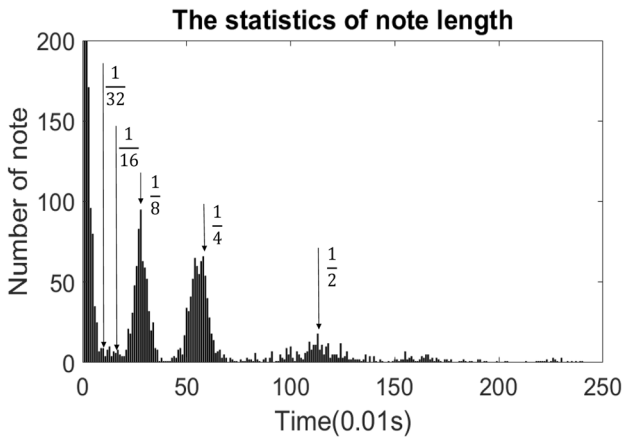

Fig. 4 The relationship between time (note type) and note appearance number extracted from the BACH10 database

Pitch Domain Analysis

After that, a matrix of pitch candidates and their corresponding harmonics ***PCH***$\epsilon\mathbb{R}^{N*H}$ can be extended from ***PD***(n). The first column of this matrix is non-zero pitch values and the rest of the columns have the associated harmonic pitches of each non-zero pitch, where the harmonic pitch is the corresponding pitch value of the harmonic frequency. A harmonic map ***HMap***$\epsilon\mathbb{R}^{M*H}$ is employed here to guide the extension process, which includes the pianoroll number (m) of the fundamental frequency ($\boldsymbol{F}_0$) and the corresponding harmonic frequency for every note. Following the MIDI tuning standard, we transfer the nth non-zero fundamental frequency to its corresponding pianoroll number using Eq. (12). Here, ***PD*** needs to be subtracted by 20 due to the difference between the pianoroll and the MIDI number:

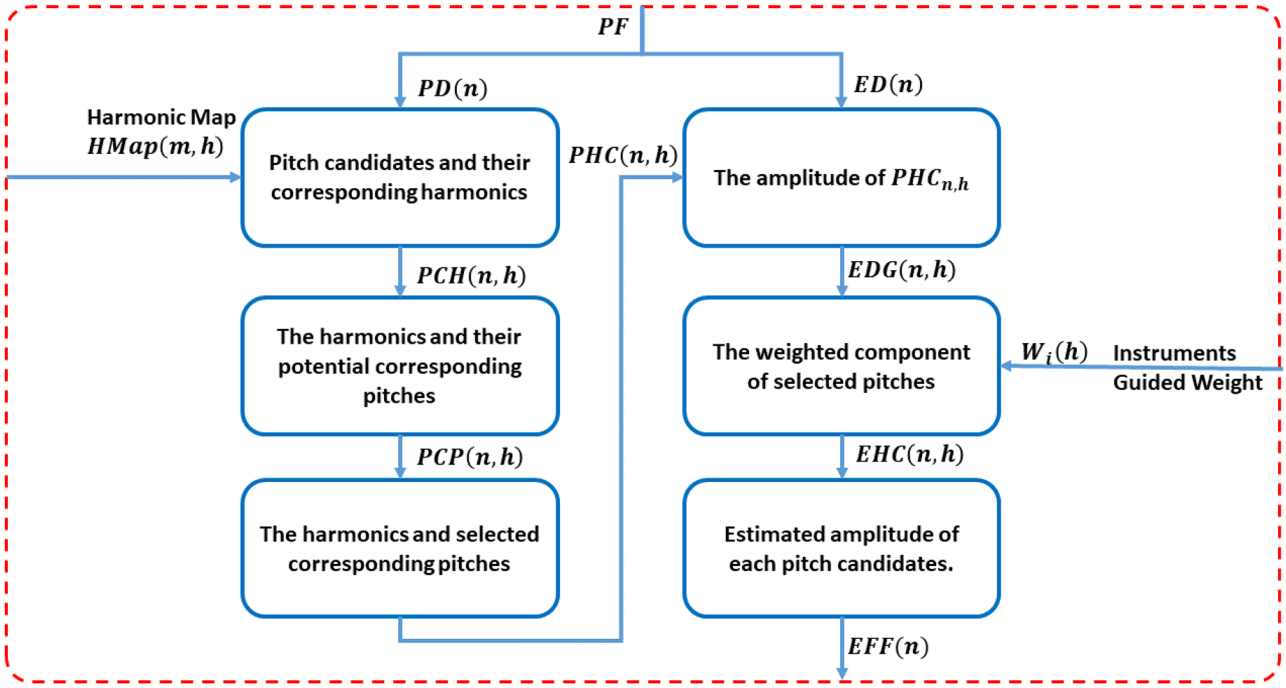

Fig. 5 Flowchart of the proposed HSD

$$\begin{aligned} \boldsymbol{PD}(n) &= 69 + 12\log_2\left(\frac{\boldsymbol{F}_0(n)}{440Hz}\right) \\ \boldsymbol{m}(n) &= \boldsymbol{PD}(n) - 20,\ |\boldsymbol{m}|\epsilon[1, 88] \end{aligned} \tag{12}$$

where 69 and 440 are the values of the MIDI number and frequency for the standard A, respectively. Twelve is the number of notes in one octave. Given a frequency of the input signal, its harmonic frequencies are multiples of the fundamental frequency. In this study, we set concert A as 440 Hz for fast implementation. Note that the concert A is not always the standard A with 440 Hz, where its frequency may vary depending on the playing style of the instruments and music pieces. It is worth mentioning that our algorithm does not reply on the frequency setting of concert A, as our algorithm focuses on the analysis of the relationship between fundamental frequency and harmonic frequencies, which mainly depends on the music temperament.

An example of calculating MIDI number of harmonic frequency in ***HMap*** is given in Table 2.

PCH(n, h) is the h^{th} harmonic pitch component of the pitch n where n lies within $[1, N]$ and h is within $[1, H]$. H is set to 5 in the experiment, and N is the number of non-zero value in each frame:

Table 1 Description of parameters

Parameters	Definition	Index/Dimension
N	The number of non-zero fundamental-pitch;	$n\epsilon[1, N]$
H	The number of harmonic-pitch; default is 5	$h\epsilon[1, H]$
I	The number of the instruments in the music piece	$i\epsilon[1, I]$
m	Vector of pianoroll	$\mathbb{R}^{N*1}$
PF	Spectrogram of SI-PLCA after filtering	$\mathbb{R}^{88*Time}$
PD	Pitch value of PF	$\mathbb{R}^{N*1}$
PCH	Value of pitch candidates and their corresponding harmonics	$\mathbb{R}^{N*H}$
PCP	Value of harmonics and their potential corresponding pitches	$\mathbb{R}^{N*H}$
PHC	Value of harmonics and selected pitches	$\mathbb{R}^{N*H}$
ED	Energy value of PF	$\mathbb{R}^{N*1}$
EDG	Amplitude of fundamental pitch and their corresponding harmonic	$\mathbb{R}^{N*H}$
EHC	Amplitude of harmonic components presented in the pitch n	$\mathbb{R}^{N*H}$
EFF	Final result of pitch amplitude	$\mathbb{R}^{N*1}$

Table 2 Example of calculating A4 in the ***HMap***

Attribute	Fundamental frequency,F_0	Harmonic Frequency, $k \times F_0$ (Hz)			
		$2\,F_0$	$3\,F_0$	$4\,F_0$	$5\,F_0$
Frequency (Hz)	440	880	1320	1760	2200
Pianoroll	49	61	68	73	77
MIDI number	69	81	88	93	97
Letter name	A4	A5	E6	A7	C#7/Db7

$$\boldsymbol{PCH}(n,h) = \boldsymbol{HMap}(\boldsymbol{m}(n),h),\ \boldsymbol{PCH}\epsilon\mathbb{R}^{N\times H} \tag{13}$$

Let ***PCP*** be a matrix of the harmonics and their potential corresponding pitches, which contains the harmonic components and their associated pitches being calculated from the original pitch at a specific value of h as follows:

$$\delta(x-y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

$$\boldsymbol{PCP}(n,h) = \boldsymbol{PCH}(n,h)\cdot\delta[\boldsymbol{PCH}(n,h) - \boldsymbol{PCH}(n,1)],\ \boldsymbol{PCP}\epsilon\mathbb{R}^{N\times H} \tag{15}$$

where $\delta(x-y)$ is a function of the equivalence gate with two inputs. The output of the equivalence gate will be 1 if the two inputs equals (i.e. $h = 1$). Otherwise, it will become zero. Using Eqs. (14) and (15), $\boldsymbol{PCP}(n,h)$ can be identified for each harmonic component.

Let $\boldsymbol{PHC}(n,1)$ be a harmonic component, and $\boldsymbol{PHC}(n,h)(h = 2,\ldots,H)$ represents the relative associated pitches. ***PHC*** is the value that correlates to ***PCP*** in identifying potentially the original pitch values. The matrix for all of the potentially original pitch values is estimated below. If $\boldsymbol{PCP}(n,h) = \boldsymbol{PCP}(n,1)$, an equivalence gate value of 1 is assigned, and the output value from the square brackets becomes 1 in Eq. (16):

$$\begin{aligned}\boldsymbol{PHC}(n,h) &= \boldsymbol{PCP}(n,1)\cdot\delta[\boldsymbol{PCP}(n,h) - \boldsymbol{PCP}(n,1)],\\ \boldsymbol{PHC}\ &\epsilon\ \mathbb{R}^{N\times H},\ n\ \epsilon\ [1,N],\ h\ \epsilon\ [1,H]\end{aligned} \tag{16}$$

Energy Domain Analysis

In the energy domain, $\boldsymbol{EDG}(n,h)$ is a value generated from $\boldsymbol{ED}\epsilon\mathbb{R}^{N*H}$ and $\boldsymbol{PHC}(n,h)$ as defined below:

$$\boldsymbol{EDG}(n,h) = \boldsymbol{ED}(n)\cdot\delta[\boldsymbol{PHC}(n,h) - \boldsymbol{PHC}(n,1)],\ \boldsymbol{EDG}\epsilon\mathbb{R}^{N\times H} \tag{17}$$

In the following, we will describe two cognitive theories which have inspired our proposed guided weight mechanism for fundamental frequency detection. First, according to the harmonic periodicity and instrument timbre theory [18], the harmonic periodicity of different instruments should be the same, although the sound of which varies by their timbres as reflected on the ratio of harmonic amplitude to the fundamental amplitude [35]. The instruments from different families will have a large ratio, and vice versa. For the instrument that produces a sound from strings such as piano, and violin (Fig. 6d), their harmonic amplitudes generally decrease gradually. On a different note, for woodwind instruments such as clarinet (Fig. 6c) and bassoon (Fig. 6a), the amplitudes of their first harmonic would be lower than that of their second harmonic. Therefore, the energy ratio of the fundamental frequency and harmonic frequency energy (timbre) is unaffected by monophonic or polyphonic textures, but unique in individual instruments. Second, according to the acoustic theory [36], when two or more sound waves occupy the same space, they move through rather than bounce off each other. For example, the result of any combination of sound waves is simply the addition of these waves. Theoretically, the energy of the mixed monophonic and polyphonic audio should be the same, though there is unavoidable difference in the real case. The results of a single frame after step 1 (section III-B) of the harmonic structure detection (HSD) are plotted as profile of pitch values as shown in Fig. 6. The profiles of four single music sources are shown in Fig. 6a–d. The profile of the mixed monophonic notes is given in Fig. 6e, which is composed of four single music sources, i.e. notes no. 1–no. 4, and the profile of the polyphonic notes shown in Fig. 6f is generated from one mixed channel. Considering that the profile of mixed monophonic notes is the ideal value, and the profile of the polyphonic notes is the predicted actual value. As seen in Fig. 6f, there are few amplitude differences between the profiles of the polyphonic and monophonic notes due to the resonance in the polyphonic notes and channel distortion during data recording and transmission, but the overall trend of the two profiles is very similar.

Motivated by these, we proposed the guided weight mechanism which is denoted as Eq. (18) in our model for improving the detection of the fundamental frequency. The guiding weight is calculated by the averaged ratio of the amplitude of harmonic $\boldsymbol{ED\_mono}(h)$ and fundamental frequency $\boldsymbol{ED\_mono}(1)$ in the monophonic data, before applying to the polyphonic data. The variable I is the number of known instruments that can be identified in the music piece:

$$\boldsymbol{W}_i(h) = \frac{1}{T}\sum_{t=1}^{T}\frac{\boldsymbol{ED\_mono}_t(h)}{\boldsymbol{ED\_mono}_t(1)},\ h\ \epsilon\ [1,H],\ i\ \epsilon\ [1,I] \tag{18}$$

where T is the number of time frames in the monophonic data, the first non-zero value of $\boldsymbol{ED\_mono}_t(1)$ is always the fundamental frequency, and the remaining non-zero values $\boldsymbol{ED\_mono}_t(h)$ are the harmonic frequencies.

Equation (19) estimates the amplitude of harmonic components (***EHC***) presented in the pitch n by multiplying the guided weight of selected instrument with ***EDG***.

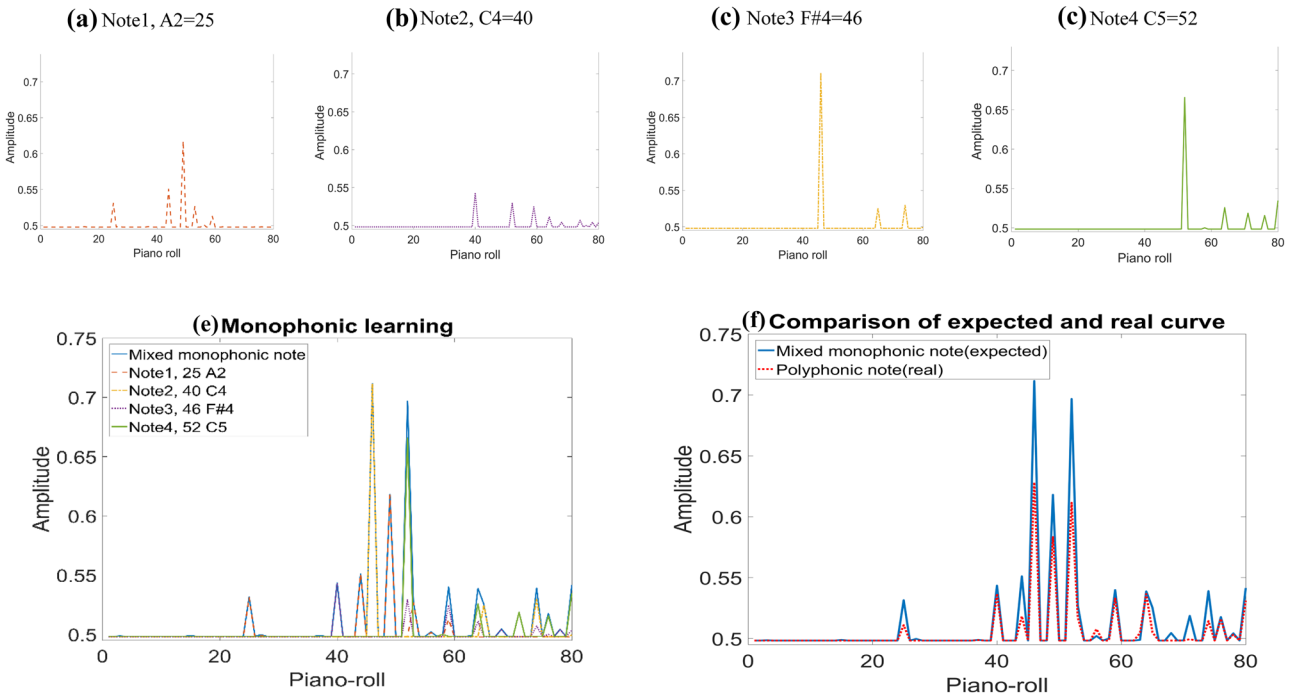

Fig. 6 **a** Note1, A2 = 25. **b** Note2, C4 = 40. **c** Note3 F#4 = 46. **d** Note4 C5 = 52. Profile of pitch values for monophonic and polyphonic analysis in a single frame. Single notes with its MIDI number for bassoon (**a**), saxophone (**b**), clarinet (**c**) and violin (**d**); **e** is the monophonic learning result when combining the four note; **f** is the comparison of real polyphonic value with expected mixed monophonic notes

Theoretically, the amplitude of harmonic should be a portion to the amplitude of the fundamental frequencies. It is noted that the fundamental frequencies must occur at $h = 1$, then harmonic frequencies occur at $h = 2{:}H$.

$$\begin{aligned} &\boldsymbol{EHC}_i(n,h) = \boldsymbol{EDG}(n,h) \cdot \boldsymbol{W}_i(h), \\ &\boldsymbol{EHC}_i \epsilon \mathbb{R}^{N*H},\ n\ \epsilon\ [1,N],\ h\ \epsilon\ [1,H] \end{aligned} \tag{19}$$

Based on the $\boldsymbol{EHC}_i$ determined from Eq. (19), the amplitude of fundamental frequency in pitch n after subtracting the summed harmonic components' amplitude will be kept updating until the fundamental frequencies from all instruments are estimated.

$$\boldsymbol{ED}(n) = \boldsymbol{EHC}_i(n,1) - \sum\nolimits_{h=2}^{H} \boldsymbol{EHC}_i(n,h) \tag{20}$$

Eventually, the amplitude of fundamental frequency in pitch n, represented as $\boldsymbol{EFF}$, can be obtained by Eq. (21).

$$\boldsymbol{EFF}(n) = \boldsymbol{ED}(n),\ \boldsymbol{EFF} \epsilon \mathbb{R}^{N\times 1} \tag{21}$$

For each non-zero pitch n in each frame t, it will have a rank value $\boldsymbol{R}(n)$ according to the $\boldsymbol{EFF}(n)$, then a 2D rank map $R(n,t)$ will be generated for the whole music piece, i.e. pitch/pianoroll vs. time frame as shown in Fig. 3b, which will be used to fully represent our detected harmonic structure. A brief implementation of energy domain procedure is summarized in Algorithm 1.

Algorithm 1

Inputs:$\boldsymbol{ED}(\boldsymbol{n})$

Step 1: Generate a matrix including the amplitude of fundamental pitch and their corresponding harmonic pitches using Eq. (17)

Step 2: Calculate the weight for each type of instrument using Eq. (18)

Step 3: Estimate the amplitude of harmonic components ($\boldsymbol{EHC}$) presented in the pitch n using Eq. (19)

Step 4: Update $\boldsymbol{ED}$ by Eq. (20)

Step 5: Repeat steps 1–4 until the fundamental frequencies from all instruments are estimated

Obtain the final estimated amplitude of fundamental frequency in pitch n by Eq. (21)

Note Tracking

As seen in Fig. 3b, although most fundamental pitches have been extracted, the notes still show a poor consistency. To improve this, a note tracking method based on the music perception and multi-pitch probability weight was proposed. According to the music theory [33], the occurrence

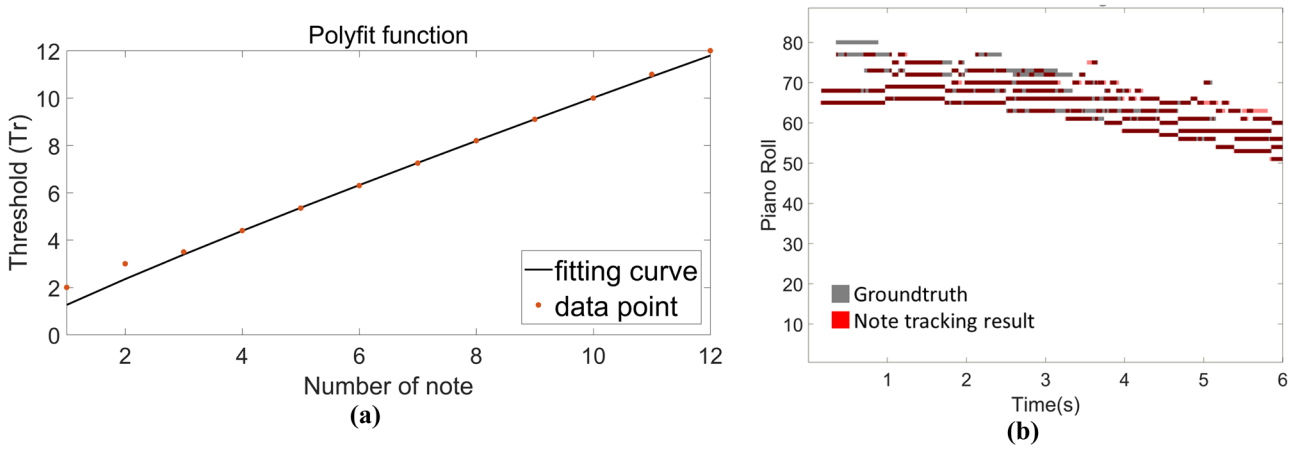

Fig. 7 Poly function of threshold $\boldsymbol{TH}_2$ (**a**) and results from the note tracking in comparison to the ground truth (**b**)

of demisemiquaver is generally quite low in music pieces. As a result, notes with a length shorter than demisemiquaver are filtered out. The averaged rank of the connected pitch group in the rank map is calculated and denoted as $\overline{R}$. If $\overline{R}$ is larger than an adaptive threshold TH_2, the pitch group is considered a harmonic and will be skipped from the analysis. As the polyphonic music pitches vary over time, the TH_2 will also change accordingly. To account for this change, a fitting function was generated for TH_2 (Fig. 7a), which is adaptive to the number of notes $x \in [1, 12]$ for each frame, as given:

$$TH_2 = 1.26x^{0.9} \tag{22}$$

The fitting curve of TH_2 is obtained by minimising the fitting error between ground truth and our estimate. Figure 7b displays the note tracking results where most of the noise and the inconsistencies have been filtered out. The result has also achieved a similar profile to that of the ground truth data.

Experimental Results

Experimental Settings

To validate the effectiveness of the proposed approach, the first dataset used for evaluation is the MIDI Aligned Piano Sounds (MAPS) [37], in which all music pieces are recorded in the MIDI format initially and then converted into ".wav" format. MAPS also have differently purposed subsets such as monophonic excerpts and chords. For this case, only one subset is used which includes polyphonic music pieces. In addition, there are several instruments and recording conditions in MAPS. The "ENSTDkCI" is chosen as the music played using a real piano rather than an acoustic one, i.e. a virtual instrument, and recording occurs in soundproofed conditions. The second dataset is BACH10 [38], which contains 10 pieces using violin, clarinet, saxophone and bassoon from J.S.Bach chorales, where each piece lasts approximately 30 s. The third dataset is TRIOS [39], which is the most complex one among the three as it contains five multitrack chamber music trio pieces. The sampling rate for all music pieces is 44,100 Hz.

For objective assessment, the most commonly used frame-based metric, F-measure (F_1) [40, 41], is adopted. It combines the positive predictive value (PPV, also namely precision) and the true positive rate (TPR, also namely recall) for a comprehensive evaluation as follows:

$$F_1 = \frac{2 \cdot PPV \cdot TPR}{PPV + TPR} \tag{23}$$

where $TPR = \frac{T_p}{T_p+F_n}$, $PPV = \frac{T_p}{T_p+F_p}$, and T_p, F_p and F_n refer respectively to the number of correctly detected F_0, incorrectly detected F_0 and missing detection of the F_0. Specifically, these three components can be calculated by comparing the binary masks of the detected MPE results and the ground truth.

Performance Evaluation

Table 3 shows the quantitative assessment of 11 benchmarking methods on MAPS, BACH10 and TRIOS datasets. We divide all benchmarking methods into two categories: shallow learning method and DL method. Weak learning methods include a traditional machine learning model or a prior knowledge-based model whereas DL methods include deep neural networks and deep convolutional neural networks.

Many MPE approaches select a pair of methods from CQT, PLCA, equivalent rectangular bandwidth (ERB) and NMF for pianoroll transcription. Therefore, two of the most representative methods, i.e. CQT + PLCA proposed

Table 3 Frame-level performance of different methods on three datasets

Category	Methods	F_1		
		MAPS	BACH10	TRIOS
Shallow learning	Benetos [43]	64.17	68.40	***66.46***
	Benetos [31]	59.31	70.57	64.93
	Vincent [42]	***72.35***	79.78	59.40
	Duan [38]	67.41	70.90	45.80
	Klapuri [3]	60.10	68.30	50.50
	CFP [8]	68.67	**85.51**	64.64
	SONIC [44]	63.60	66.49	56.65
	HSD(proposed)	**76.30**	***80.17***	**67.63**
Deep learning	ConvNet [15]	64.14	–	–
	RNN [15]	57.67	–	–
	Li [40]	69.42	–	–
	INN [41]	72.29	–	–

Top two methods are bold with the second also italic

by Benotos and Dixon [31] and ERB + NMF proposed by Vincent et al. [42], are chosen for benchmarking. In Table 3, Benetos et al. [43] and Vincent [42] can produce the second best performance on the MAPS and TRIOS datasets, respectively, which validates the effectiveness of CQT + PLCA and ERB + NMF. However, due to the lack of efficient harmonic analysis, the performance of both methods is inferior to the proposed HSD method. Unlike the methods from Benetos and Vincent, other methods adopt different ideas for MPE. SONIC [44] proposed a connectionist approach where an adaptive oscillator network was used to track the partials in the music signal. However, without a matrix factorization process, its performance is limited on the three datasets. Su and Yang [8] proposed a combined frequency and periodicity (CFP) method to detect the pitch in both frequency domain and lag (frequency) domain. The CFP method in Table 3 gives the best performance on the BACH10 dataset, but relatively poorer results on the other two datasets. The main reason here is possibly because the music pieces in the MAPS and TRIOS datasets have more short notes than those in the BACH10 dataset, and CFP has the limited ability for detecting the short notes but exhibit less errors for continuous long notes. Furthermore, the assumption of CFP does not hold for high-pitch notes of piano, as both MAPS and TRIOS have many piano music pieces. In addition, the music pieces in the MAPS database contain multiple notes in most frames, which have led to extra difficulty for polyphonic detection. However, the proposed method can still successfully solve this problem by effectively analysing the relationship of the position and energy between the fundamental frequency and harmonic frequencies for the notes. As a result, the performance of the proposed method on MAPS is the best, which is 8% higher than that of CFP. Klapuri [3] proposed an auditory model–based F_0 estimator, and Duan [38] proposed a maximum-likelihood approach for multiple F_0 estimation, but both methods result in inferior performance compared to the results achieved by Benetos et al. [31, 43], Vincent et al. [42] or CFP [8]. Furthermore, Klapuri's [3] and Duan et al.'s [38] methods lack an effective pre-processing stage (i.e. TF representation and matrix factorization) or harmonic analysis, which is the main reason why their overall performance is less effective in comparison to ours.

The proposed method was also compared with four deep learning–based supervised approaches on MAPS dataset. Due to lack of publicly available source codes, only the data that was reported in the original paper was duplicated for comparison. The first two methods are proposed by Sigtia et al. [15], which are mainly based on the music language models (MLMs). However, due to insufficiently labelled data in the existing polyphonic music databases for training, such limitations have affected further analysis of DL-based approaches. Furthermore, the MLM model is not robust to ambient noise, whereas music pieces in reality generally contain a lot of ambient noise. This has resulted in DL-based methods failing to fully analyse the inner structure of the music pieces. As a result, DL-based methods cannot achieve the same performance as the HSD method or some of the other unsupervised methods such as Benetos et al. [43] on the MAPS dataset. Su [40] and Kelz [41] also proposed DL-based methods for AMT. Although better than [15], their performance is still not ideal as there is insufficient music knowledge support embedded. To this end, more music theories should be introduced for improved AMT.

In summary, referring to Table 3, the proposed method yields the best results on both the MAPS and TRIOS datasets, also the second-best in BACH10 according to F_1 value, thanks to the guidance of music cognition. However, the method can still be improved, especially for reducing the computation cost. As it takes 2 min to process a 30-s music piece, this is longer than some other methods. In addition, although the profile of the real polyphonic note is close to the expected mixed monophonic note, as shown in Fig. 6e, f, there are still some differences in the final values of the monophonic and polyphonic profiles which can be further improved.

Key Stage Analysis

In this section, the contribution of several major stages in the proposed MPE system is discussed, where the performance of each stage is evaluated on the MAPS dataset in terms of the precision, recall and F_1. To calculate these three metrics, the result of each stage is normalised by using Eqs. (8) and (9), and the results are binarized with a fixed threshold value of 0.5.

Table 4 System configuration

Configurations	Precision	Recall	F-measure
A	0.408	0.879	0.545
A + B	0.438	0.876	0.571
A + B + C	0.747	0.718	0.725
A + B + C + D	0.753	0.773	**0.763**

The bold one indicates the best performed result in the column

We generalize our proposed MPE system into four key stages detailed as follows:

- Stage A: The transcription map from SI-PLCA and CQT.
- Stage B: The result after applying the first-step HSD.
- Stage C: The result after applying the second-step HSD.
- Stage D: The result after applying note tracking.

Table 4 illustrates the details of the system configurations. By combination of different key stages, the corresponding system is built up for evaluation. Each stage has specific components which are indispensable to the results of the system. Stage A shows the highest recall and lowest precision after applying CQT and SI-PLCA. The presence of F_0 and harmonics is all detected; however, many amplitudes are concentrated in higher frequency (harmonic) regions which inhibits the identification of F_0. After combining stage B, the recall value decreases by 0.03%, but the precision value increases by almost 3%. This is mainly due to the removal of noise in HSD. In stage C, the core of the MPE system contributes to an increase of nearly 30% for precision and 15–18% for F_1 compared to previous combinations. Finally, after applying the proposed note tracking step (stage D), the recall value is further improved by 5.5% which leads to the final F_1 value improved by 3.8% compared to the previous stage.

Table 5 Time-freqeuncy transform and piano-roll transcription comparison

Methods	AUC	MAE	maxF
ERB + PLCA	0.922	0.0403	0.6687
ERB + CNMF	0.939	0.0487	**0.7213**
CQT + PLCA	**0.942**	**0.0390**	0.7089
CQT + CNMF	0.906	0.0411	0.6296

The bold one indicates the best performed result in the column

Assessment of CQT and ERB

In our proposed MPE system, CQT is employed to model the human cochlea perception. However, cochlea perception is not always constant in Q. Therefore, apart from CQT, the equivalent rectangular bandwidth (ERB) method is also widely used for time–frequency transform [42]. As most ERB methods are actually based on the Gamma tone filter-bank to model the human auditory system [45], it decomposes a signal and passes it through a bank of gamma tone filters, equally spaced on the equivalent rectangular bandwidth (ERB) scale. However, ERB methods may not be necessary to produce better MPE performance than CQT. To further validate this assumption, we have combined CQT [27] and ERB [42] pair-wisely with PLCA [43] and NMF [42] to form four hybrid methods, i.e. CQT + PLCA, CQT + NMF, ERB + PLCA and ERB + NMF, for quantitative analysis in terms of the precision-recall, ROC, F-measure curve (Fig. 8), AUC, MAE and maxF (Table 5). Here AUC, MAE, and maxF denote respectively the area under the ROC curve, the mean average error and the max value of F-measure curve. These three criteria have the same importance. As seen in Fig. 8, the ERB + NMF and CQT + PLCA show comparable results; both outperform the other two methods. In Table 5, although ERB + NMF gives the best maxF value, CQT + PLCA gives the best AUC and lowest MAE, indicating a smaller false alarm. Therefore, CQT + PLCA is the best among these four

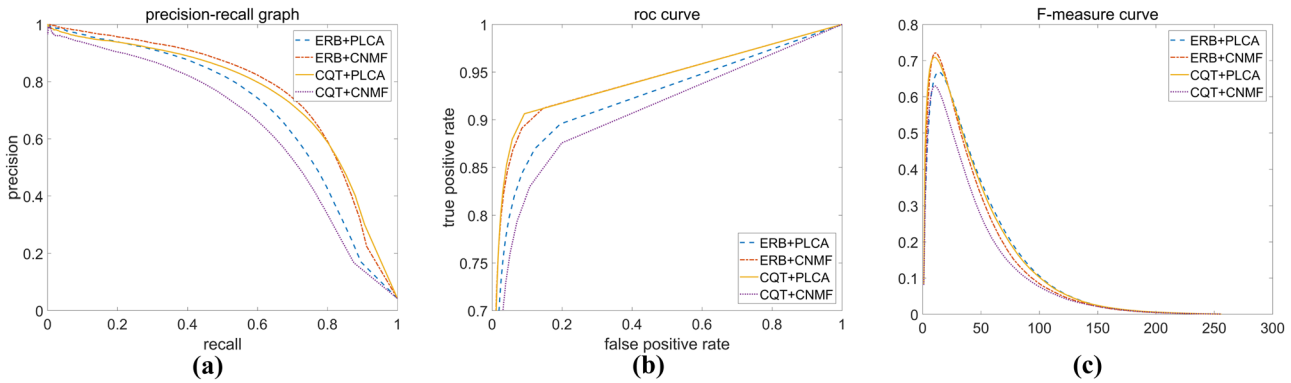

Fig. 8 Precision-recall, receiver operating characteristic and F-measure curve of four pair-wise methods where pairwise combinations of CQT and ERB with PLCA and NMF, respectively, are compared

methods, which is also the main reason why it is used in our proposed MPE system.

Conclusion

In this paper, a harmonic analysis method is proposed for the MPE system, inspired by music cognition and perception. CQT and SI-PLCA are employed in the pre-processing stage for pianoroll transcription in mixture music audio signal, from which the proposed HSD is used to extract the multi-pitch pianorolls. The proposed MPE system is not limited by the number of instruments. For multi-instrument cases (i.e. symphony in BACH10 and TRIOS datasets), the mixture characteristics of each instrument can be extracted for adaptive detection of the fundamental frequencies. From the experiment results, the proposed MPE system yields the best performance on the MAPS and TRIOS datasets, and the second-best on the BACH10 dataset. Through investigation of the performance of key components, the HSD provided the greatest contribution to the system, which validates the value of adding an efficient harmonic analysis model for improving significantly the performance of the MPE system. Furthermore, adding note tracking can further improve the efficacy of the MPE system.

However, the proposed MPE system still has much room to improve. First, it is worth mentioning that the expectation maximization (EM) algorithm has some limitations, especially the low convergence speed, sensitive to initial settings and inherent non-convex caused local optimum. As a result, it makes PLCA very time consuming, even unsuitable for processing large datasets. Therefore, how to better select the initial value and speed up the convergence can be a valuable work for future investigation. Second, the assumption of knowing the type of instruments in the music pieces is often unrealistic in real scenarios. Therefore, blind source separation can be integrated in our model to tackle this limitation. Third, analysis of the beat and chord along with integrated deep-learning models such as transformer networks [46] and long-short term memory [47] can be considered to further enhance the accuracy of pitch estimation. On the other hand, introducing more music perceptions such as ornaments and rhythm into the model will be helpful for more precise interpreting of the music pieces. Furthermore, an improved note tracking process can be introduced by fusing self-attention [48] and natural language processing model [49]. Finally, testing on larger datasets such as MusicNet [50] and MAESTRO [51] will be beneficial for more comprehensive modelling and validation.

Acknowledgements This work is partially supported by the National Nature Science Foundation of China (grant 61876125) and the University of Strathclyde PhD Studentship.

Declarations

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of interest The authors declare no competing interests.

References

1. Benetos E, Dixon S, Duan Z, Ewert S. Automatic music transcription: an overview. IEEE Signal Process Mag. 2018;36(1):20–30.
2. Emiya V, Badeau R, David B. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Trans Audio Speech Lang Process. 2010;18(6):1643–54.
3. Klapuri A. Multipitch analysis of polyphonic music and speech signals using an auditory model. IEEE Trans Audio Speech Lang Process. 2008;16(2):255–66.
4. Bay M, A Ehmann F, Downie JS. Evaluation of multiple-F0 estimation and tracking systems. In: ISMIR; 2009. pp 315–20.
5. Benetos E, Dixon S, Giannoulis D, Kirchhoff H, Klapuri A. Automatic music transcription: challenges and future directions. J Intel Inf Syst. 2013;41(3):407–34.
6. Chunghsin Y. Multiple fundamental frequency estimation of polyphonic recordings. University Paris 6; 2008. Ph. D. dissertation.
7. Benetos E, Dixon S. Joint multi-pitch detection using harmonic envelope estimation for polyphonic music transcription. IEEE Journal of Selected Topics in Signal Processing. 2011;5(6):1111–23.
8. Su L, Yang Y-H. Combining spectral and temporal representations for multipitch estimation of polyphonic music, *IEEE/ACM Transactions on Audio*. Speech and Language Processing (TASLP). 2015;23(10):1600–12.
9. Fuentes B, Badeau R, Richard G. Adaptive harmonic time-frequency decomposition of audio using shift-invariant PLCA. In: Proc. ICASSP; 2011. p. 401–4.
10. Vincent E, Plumbley MD. Efficient Bayesian inference for harmonic models via adaptive posterior factorization. Neurocomputing. 2008;72(1–3):79–87.
11. Cheuk KW, Luo Y-J, Benetos E, Herremans D. The effect of spectrogram reconstruction on automatic music transcription: an alternative approach to improve transcription accuracy. In: Proc. ICPR; 2021. p. 9091–8.
12. Mukherjee H, Obaidullah SM, Phadikar S, Roy K. MISNA-a musical instrument segregation system from noisy audio with LPCC-S features and extreme learning. Multimedia Tools and Applications. 2018;77(21):27997–8022.
13. Mukherjee H, Dhar A, Obaidullah SM, Santosh K, Phadikar S, Roy K. Segregating musical chords for automatic music transcription: a LSTM-RNN approach. In: International Conference on

Pattern Recognition and Machine Intelligence. Springer; 2019. p. 427–35.
14. Fan Z-C, Jang J-SR, Lu C-L. Singing voice separation and pitch extraction from monaural polyphonic audio music via DNN and adaptive pitch tracking. In: Proc. Multimedia Big Data (BigMM); 2016. p. 178–85.
15. Sigtia S, Benetos E, Dixon S. An end-to-end neural network for polyphonic piano music transcription. IEEE/ACM Transactions on Audio Speech and Language Processing (TASLP). 2016;24(5):927–39.
16. Yan Y, et al. Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement. Pattern Recogn. 2018;79:65–78.
17. Pichevar R, Rouat J. Monophonic sound source separation with an unsupervised network of spiking neurones. Neurocomputing. 2007;71(1–3):109–20.
18. Fletcher NH, Rossing TD. The physics of musical instruments. Springer Science & Business Media; 2012.
19. Justus TC, Bharucha JJ. Music perception and cognition. In: Stevens' Handbook of Experimental Psychology, Sensation and Perception. John Wiley & Sons Inc; 2002. p. 453.
20. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B. 1977;39:1–38.
21. Bernardo JM, Smith AF. Bayesian theory. IOP Publishing; 2001.
22. Emiya V, Badeau R, David B. Multipitch estimation of quasi-harmonic sounds in colored noise. In: 10th Int. Conf. on Digital Audio Effects (DAFx-07); 2007.
23. Duan Z, Temperley D. Note-level music transcription by maximum likelihood sampling. In: ISMIR; 2014. p. 181–6.
24. Alvarado Duran PA. Acoustically inspired probabilistic time-domain music transcription and source separation. Queen Mary University of London; 2020.
25. Nishikimi R, Nakamura E, Itoyama K, Yoshii K. Musical note estimation for F0 trajectories of singing voices based on a Bayesian semi-beat-synchronous HMM. In: ISMIR; 2016. p. 461–7.
26. Gowrishankar BS, Bhajantri NU. An exhaustive review of automatic music transcription techniques: survey of music transcription techniques. In: Proc. Signal Processing, Communication, Power and Embedded System; 2016. p. 140–52.
27. Brown JC. Calculation of a constant Q spectral transform. The Journal of the Acoustical Society of America. 1991;89(1):425–34.
28. Bendor D, Wang X. The neuronal representation of pitch in primate auditory cortex. Nature. 2005;436(7054):1161–5.
29. Schörkhuber C, Klapuri A. Constant-Q transform toolbox for music processing. In: 7th Sound and Music Computing Conference, Barcelona, Spain; 2010. p. 3–64.
30. Smaragdis P, Brown JC. Non-negative matrix factorization for polyphonic music transcription. In: Proc. Applications of Signal Processing to Audio and Acoustics; 2003. p. 177–80.
31. Benetos E, Dixon S. A shift-invariant latent variable model for automatic music transcription. Comput Music J. 2012;36(4):81–94.
32. Han J, Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning. In: Proc. Artificial Neural Networks; 1995. p. 195–201.
33. Smith LM. A multiresolution time-frequency analysis and interpretation of musical rhythm. Australia: University of Western Australia Perth; 2000.
34. d'Alessandro C, Castellengo M. The pitch of short-duration vibrato tones. The Journal of the Acoustical Society of America. 1994;95(3):1617–30.
35. Li X, Wang K, Soraghan J, Ren J. Fusion of Hilbert-Huang transform and deep convolutional neural network for predominant musical instruments recognition. In: Proc. Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar); 2020. p. 80–9.
36. Kinsler LE, Frey AR, Coppens AB, Sanders JV. Fundamentals of acoustics. John Wiley & Sons; 2000.
37. Emiya V, Bertin N, David B, Badeau R. MAPS-A piano database for multipitch estimation and automatic transcription of music. Research Report; 2010. p. 11. inria00544155. https://hal.inria.fr/inria-00544155/document. Accessed 12 Aug 2021.
38. 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(8):2121–33.
39. Fritsch J, Plumbley MD. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In: Proc. Acoustics, Speech and Signal Processing (ICASSP); 2013. p. 888–91.
40. Su L. Between homomorphic signal processing and deep neural networks: constructing deep algorithms for polyphonic music transcription. In: Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC); 2017. p. 884–91.
41. Kelz R, Widmer G. Towards interpretable polyphonic transcription with invertible neural networks. arXiv preprint; 2019. http://arxiv.org/abs/1909.01622. Accessed 12 Aug 2021.
42. Vincent E, Bertin N, Badeau R. Adaptive harmonic spectral decomposition for multiple pitch estimation. IEEE Trans Audio Speech Lang Process. 2010;18(3):528–37.
43. Benetos E, Cherla S, Weyde T. An efficient shift-invariant model for polyphonic music transcription. In: 6th International Workshop on Machine Learning and Music; 2013.
44. Marolt M. A connectionist approach to automatic transcription of polyphonic piano music. IEEE Trans Multimedia. 2004;6(3):439–49.
45. Smith JO, Abel JS. Bark and ERB bilinear transforms. IEEE Transactions on speech and Audio Processing. 1999;7(6):697–708.
46. Vaswani A, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.
47. Chen N, Wang S. High-level music descriptor extraction algorithm based on combination of multi-channel CNNs and LSTM. In: ISMIR; 2017. p. 509–14.
48. Parmar N, et al. Image transformer. In: International Conference on Machine Learning. PMLR; 2018. p. 4055–64.
49. Turc I, Chang M-W, Lee K, Toutanova K. Well-read students learn better: on the importance of pre-training compact models. arXiv preprint; 2019. http://arxiv.org/abs/1908.08962. Accessed 12 Aug 2021.
50. Draguns A, Ozoliņš E, Šostaks A, Apinis M, Freivalds K. Residual shuffle-exchange networks for fast processing of long sequences. Proc AAAI Conf Artif Intell. 2021;35(8):7245–53.
51. Hawthorne C, et al. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In: International Conference on Learning Representations; 2018.