Multi-Pitch Estimation of Polyphonic Music Based on Pseudo Two-Dimensional Spectrum

Weiwei Zhang, Zhe Chen, *Senior Member, IEEE*, and Fuliang Yin

***Abstract*—Multi-pitch estimation is a fundamental and key problem in music information retrieval, but still remains challenging due to the intrinsic complexity of polyphonic music. To address this problem, a pseudo 2-D spectrum-based method is proposed in this article. The pseudo 2-D spectrum is first constructed to map the time domain signal into the 2-D frequency space, where the harmonic signal exhibits a typical 2-D pattern. Then, pitch estimation is carried out by cross-correlation between the pseudo 2-D spectrum and the fixed 2-D harmonic template. Finally, the pitches of adjacent frames are grouped into pitch contours, where the contours whose lengths are shorter than the minimum note length limitation are discarded. And the remained pitches are refined using the estimates of neighboring frames by removing probable errors and reconstructing estimates. The proposed method exploits the harmonic structure of pitched sounds in a two-dimensional frequency plane, can work in the case where some notes contain few harmonics, and the harmonic overlap proportions are reduced greatly in the harmony cases. The experimental results show that the proposed method achieves promising performance comparing with the state-of-the-art methods on the evaluation datasets, and outperforms the bispectrum-based method on both evaluation datasets.**

***Index Terms*—Multi-pitch estimation, polyphonic music transcription, pseudo two-dimensional spectrum, two-dimensional harmonic pattern.**

I. Introduction

MULTI-PITCH estimation, also known as multiple fundamental frequency estimation, refers to extracting the multiple pitches of polyphonic music from audio recordings. Most works focus on estimating multiple pitches from monaural recordings, while few tries to accomplish this task from stereophonic music mixtures [1]. Multi-pitch estimation has numerous applications, including automatic search and annotation of music, musicological analysis, auditory scene analysis, multi-instrument transcription, and so on [2]–[5].

In the past decade, a lot of effort has been made to estimate multiple pitches from polyphonic music. However, this task still remains challenging since several notes may sound simultaneously. The spectral partials of polyphonic music nearby affect each other, and even some peaks in the spectrum may relate to two or more notes due to harmony. Generally, each note in polyphonic music is characterized by its harmonic frequencies and their corresponding amplitudes, which provide the most important information for pitch estimation. But if two notes have overlapping partials, it would be difficult to assign an appropriate proportion to each note, which hinders gathering useful information for note recognition.

Since Moorer first attempted to transcribe duets [6], a lot of work has been done for polyphonic music transcription. Though some studies focus only on polyphonic music performed by piano [7], [8], a vast majority of such methods are suitable for all types of musical instruments besides piano. In [3], multi-pitch estimation methods are classified into three categories: feature-based, statistical model-based and spectrogram factorization-based methods. Typically, feature-based methods use some audio features derived from the time-frequency representation of polyphonic music to estimate multiple pitches in a joint or iterative fashion [9]–[12]. Statistical model-based methods formulate the multi-pitch estimation problem in a statistical framework, then maximum a posteriori (MAP) or maximum likelihood (ML) estimation is employed to select the most salient pitch in each iteration [4], [13]. Spectrogram factorization-based methods attempt to decompose the spectrogram of the audio mixture into a linear combination of notes with their corresponding intensities or probabilities [14]–[17]. The recordings are processed at the music-piece level, and the model parameters are often estimated using expectation maximization or non-negative matrix factorization-like algorithms.

Recently, accompanying the success of deep learning in some applications of audio signal processing, researchers also attempted to address the multi-pitch estimation problem in the deep learning framework [18]–[21]. Böck and Schedl transcribed both onsets and pitches of the notes from spectral features using bidirectional long short-term memory network [21]. Sigtia *et al.* proposed an architecture that comprises a neural network-based acoustic model and a recurrent neural network-based music language model, and reported that the acoustic model by ConvNet outperforms those by deep neural

Manuscript received October 31, 2019; revised April 29, 2020 and June 18, 2020; accepted June 21, 2020. Date of publication July 7, 2020; date of current version July 24, 2020. This work was supported in part by the National Natural Science Foundation of China under Grants 61771091 and 61871066, in part by the National High Technology Research and Development Program (863 Program) of China under Grant 2015AA016306, in part by the Natural Science Foundation of Liaoning Province under Grants 20170540159 and 20170540197, and in part by the Fundamental Research Funds for the Central Universities of China under Grants DUT17LAB04 and 3132020201. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tan Lee. *(Corresponding author: Fuliang Yin.)*

Weiwei Zhang is with the Information Science and Technology College, Dalian Maritime University, Dalian 116026, China, and also with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: zhangww@dlmu.edu.cn).

Zhe Chen and Fuliang Yin are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: zhechen@dlut.edu.cn; flyin@dlut.edu.cn).

Digital Object Identifier 10.1109/TASLP.2020.3007794

networks (DNN) and recurrent neural networks (RNN) on the MAPS dataset [18]. Instead of comparing the performance of different neural networks for the acoustic model, Kelz *et al.* focused their work on testing different input representations and techniques for training and regularization [19]. More recently, Hawthorne *et al.* developed a piano music transcription method by using a deep convolutional and recurrent neural network, where there are two branches in this system, one for note onset estimation and the other for frame-wise pitch prediction [20]. Deep learning methods are free from hand-crafted feature extraction, and their performance relies heavily on the size and diversity of datasets, which is hard to satisfy due to lack of annotated excerpts, especially for the non-piano music [18].

As far as the signal representation is concerned, most of the existing methods are based on one-dimensional spectra of different versions. The short-time Fourier transform is most commonly used [3], [9], [10], [14], [16]–[18], [21], power spectrum is also employed by several works, such as in [4] and [13], constant-Q transform (CQT) is selected in [15], and log-magnitude mel-frequency spectrogram is used in [20]. Except for the above-mentioned signal representations, Su and Yang combined the one-dimensional magnitude spectrum and temporal periodicity to detect multiple pitches [11].

Harmonic overlapping impedes note separation based on the one-dimensional spectrum, in both iterative and joint estimation fashions [3]. To address this issue, Argenti *et al.* proposed a polyphonic music transcription method based on the bispectrum [12], where the cross-correlation between the two-dimensional (2-D) harmonic template and the bispectrum is calculated to estimate the frame-wise pitches. Through this method, the overlapping harmonics from different notes are mapped into different locations in the 2-D plane. However, it is difficult to recognize the pattern if there are few harmonics or some partials are missing, and the template for the bispectrum of harmonic signals is a bit complex.

In order to reduce the proportion of harmonic overlap from concurrent notes and recognize the 2-D harmonic pattern even when there are few harmonics, a pseudo 2-D spectrum-based multi-pitch estimation method is proposed in this paper. Specifically, the pseudo 2-D spectrum is first constructed to obtain the 2-D representation of polyphonic music. Next, the preliminary pitch estimation is done by the cross-correlation between the pseudo 2-D spectrum and the 2-D harmonic template. The pitches with greater cross-correlation values are considered as potential ones. Finally, pitches are grouped to form pitch contours. The contours that do not satisfy the minimum note length limitation are discarded, and the remained pitches are refined using the estimates of neighboring frames. The proposed method exploits the harmonic structure of pitched sounds in a two-dimensional frequency plane, requires no prior information, and can work in the case where musical notes contain few harmonics. The experimental results demonstrate that the pseudo 2-D spectrum-based multi-pitch estimation method outperforms the bispectrum-based method on both evaluation datasets.

The main contributions of this paper are as follows. First, the definition and some properties of the pseudo 2-D spectrum are elaborated. Second, the 2-D pattern of harmonic signals is provided with a detailed analysis, showing why the 2-D mapping greatly reduces the proportion of harmonic overlap in the harmony cases. Finally, a multi-pitch estimation method based on the proposed pseudo 2-D spectrum is presented, the pitch estimation is carried out by cross-correlation between the pseudo 2-D spectrum and the harmonic template, and preliminary estimates are refined using the post-processing.

The organization of this paper is as follows. Section II elaborates the pseudo 2-D spectrum. The pseudo 2-D spectrum-based multi-pitch estimation method is presented in detail in Section III. The experimental results and discussions are provided in Section IV. Some conclusions are drawn in Section V. Finally, some properties of the pseudo 2-D spectrum are given in Appendix.

II. Pseudo Two-Dimensional Spectrum

In this section, the pseudo 2-D spectrum will be elaborated in detail. To illustrate its difference with the bispectrum, the definition and weakness of bispectrum for pitch estimation are first briefly provided.

A. Bispectrum

The bispectrum of signal $x(t)$ is defined as [12]

$$B_x(f_1,f_2)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}c_{3x}(\tau_1,\tau_2)\mathrm{e}^{-j2\pi(f_1\tau_1+f_2\tau_2)}\mathrm{d}\tau_1\mathrm{d}\tau_2 = X(f_1)X(f_2)X^*(f_1+f_2) \quad (1)$$

where $X(\cdot)$ and $c_{3x}(\cdot,\cdot)$ are the 1-D Fourier transform and third-order cumulant of $x(t)$, respectively, f_1 and f_2 are the two independent frequency variables in the 2-D plane, and $(\cdot)^*$ is the complex conjugate operator.

The monophonic signal containing H harmonics can be represented as

$$z(t)=\sum_{l=1}^{H}a_l\mathrm{e}^{j2\pi lf_0t} \quad (2)$$

where f_0 is the fundamental frequency, and a_l is the amplitude of the l-th harmonic.

According to the definition of Eq. (1), the bispectrum of $z(t)$ is

$$B_z(f_1,f_2)=\sum_{l=1}^{\lfloor H/2\rfloor}a_l\delta(f_1-lf_0)\sum_{m=l}^{H-l}[a_m\delta(f_2-mf_0)\cdot a_{m+l}\delta(f_1+f_2-lf_0-mf_0)] \quad (3)$$

where $\lfloor\cdot\rfloor$ denotes rounding towards negative infinity.

The amplitude of $B_z(f_1,f_2)$ is the multiplication of the m-th, l-th and $(m+l)$-th harmonics, i.e., a_m, a_l and a_{m+l}. Hence, when either of these three terms is zero, $B_z(f_1,f_2)=0$, in other words, the harmonic structure in the 2-D plane is corrupted. Moreover, even either of them is close to zero, it would be difficult to recognize the pattern. According to some previous studies, the spectral slopes of music instruments decay 3 dB to 12 dB per octave [22], [23]. Hence, the higher harmonic term, a_{m+l}, is sometimes found to be unreliable for the 2-D pattern

recognition due to its small value. To address this problem, a pseudo 2-D spectrum is proposed for multi-pitch estimation.

B. Pseudo 2-D Spectrum

The pseudo 2-D spectrum is defined as

$$P_x(f_1,f_2)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}x(t)x^*(\tau)\mathrm{e}^{-j2\pi(f_1t-f_2\tau)}\mathrm{d}t\mathrm{d}\tau$$
$$=X(f_1)X^*(f_2) \tag{4}$$

where $X(f)$ is the Fourier transform of $x(t)$.

The pseudo 2-D spectrum defined in Eq. (4) maps the one-dimensional signal into the two-dimensional plane, and can also be directly calculated using the one-dimensional spectrum.

In this paper, the pseudo 2-D spectrum is proposed for the multi-pitch estimation. In the extreme case, where $x(t)=\mathrm{e}^{j2\pi f_0t}$, $P[e^{j2\pi f_0t}]=\delta(f_1-f_0)\delta(f_2-f_0)$, i.e., even though there is only one partial in the signal, it still can be mapped into the 2-D plane, which is impossible for the bispectrum. Moreover, it can also be inferred that all harmonic signals share the same pattern. Four properties of the pseudo 2-D spectrum are listed in Appendix.

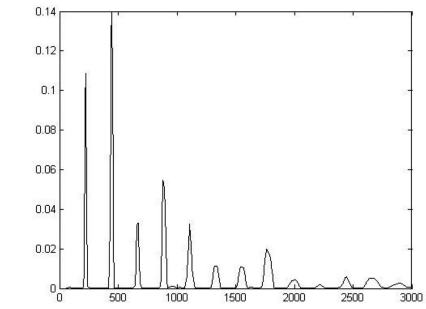
C. Pseudo 2-D Spectrum of Harmonic Signals

Pitches of polyphonic music come from notes which are harmonic signals. Hence, the pseudo 2-D spectrum of harmonic signals is discussed in this subsection. For the monophonic signal expressed in Eq. (2), its pseudo 2-D spectrum is

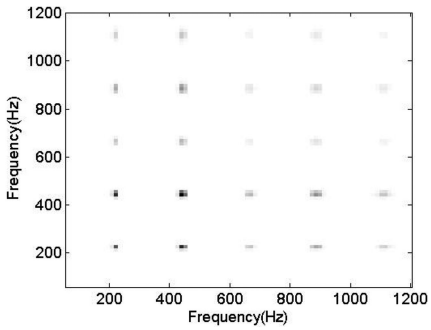
$$P_z(f_1,f_2)=\sum_{l=1}^{H}a_l\delta(f_1-lf_0)\sum_{m=1}^{H}a_m\delta(f_2-mf_0)$$
$$=\sum_{l=1}^{H}\sum_{m=1}^{H}a_la_m\delta(f_1-lf_0)\delta(f_2-mf_0). \tag{5}$$

Thus, the proposed pseudo 2-D spectrum produces an $H\times H$ pattern for a monophonic signal with H harmonics. The amplitude in the 2-D pattern is the multiplication of the l-th and m-th harmonic amplitudes, i.e., a_l and a_m. In other words, non-zero values are located at (lf_0,mf_0), where $(l,m)\in\mathbb{N}$ and $1\le(l,m)\le H$. Fig. 1 shows the spectrum and pseudo 2-D spectrum of note A3 played by a saxophone.

Suppose that the h-th harmonic is missing, the pseudo 2-D spectrum $P_z(f_1,hf_0)=0$ and $P_z(hf_0,f_2)=0$ according to Eq. (5), while the bispectrum $B_x(f_1,hf_0)=0$, $B_x(hf_0,f_2)=0$, and $B_x(lf_0,mf_0)=0$ for all combinations with $l+m=h$ according to Eq. (3). Therefore, the pseudo 2-D spectrum reduces the risk of missing the pattern compared to the bispectrum, especially for the missing fundamental and harmonic situations. Moreover, in the more common cases, where some harmonics are much smaller than others, the pseudo 2-D spectrum still performs better than the bispectrum. To illustrate the fact further, the following example is considered. The third harmonic of A3 is much smaller than the other ones, as shown in Fig. 2(a). And the bispectrum and pseudo 2-D spectrum of this note are given in Fig. 2(b) and Fig. 2(c), respectively. From Fig. 2, it can be seen that the note is easier to be recognized by the pseudo 2-D spectrum. The same phenomenon also happens when other harmonics are much smaller. This example demonstrates that the pseudo 2-D spectrum is more robust to the magnitude variation of harmonics.



(a) Spectrum



(b) Pseudo 2-D spectrum

Fig. 1. Spectrum and pseudo 2-D spectrum of A3 played by a saxophone. The pseudo 2-D spectrum is developed for multi-pitch estimation from polyphonic music in this work, and this figure is just to illustrate the difference between 1-D spectrum and the pseudo 2-D spectrum of a monophonic sound.

The polyphonic signal containing M pitched notes can be expressed as

$$z(t)=z_1(t)+\cdots+z_M(t)=\sum_{m=1}^{M}\sum_{l_m=1}^{H_m}a_{m,l_m}e^{j2\pi l_mf_{0,m}t} \tag{6}$$

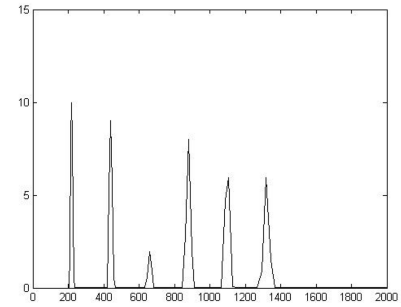
where H_m, a_{m,l_m} and $f_{0,m}$ are the harmonic number, the amplitude of the l_m-th harmonic and the fundamental frequency of note m, respectively.

The pseudo 2-D spectrum of $z(t)$ is

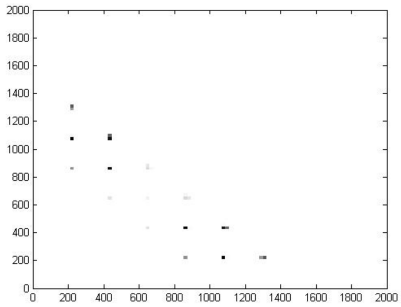
$$P_z(f_1,f_2)=\left[\sum_{m=1}^{M}\sum_{l_m=1}^{H_m}a_{m,l_m}\delta(f_1-l_mf_{0,m})\right]$$
$$\cdot\left[\sum_{n=1}^{M}\sum_{k_n=1}^{H_n}a_{n,k_n}\delta(f_2-k_nf_{0,n})\right]. \tag{7}$$

So we have

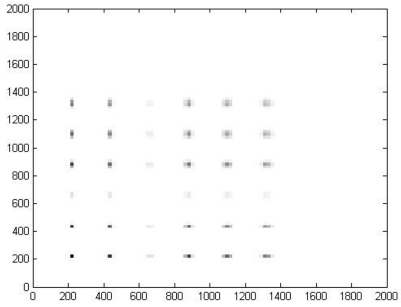
$$P_z(f_1,f_2)=\sum_{m=1}^{M}P_{z_m}(f_1,f_2)+\sum_{m=1}^{M}\sum_{\substack{n=1\\n\neq m}}^{M}C_{z_m,z_n}(f_1,f_2) \tag{8}$$



(a) Spectrum



(b) Bispectrum



(c) Pseudo 2-D spectrum

Fig. 2. Spectrum, bispectrum and pseudo 2-D spectrum of A3 whose third harmonic is much weaker than others. The pseudo 2-D spectrum is proposed for multi-pitch estimation. This figure is to illustrate the differences of 1-D spectrum, bispectrum and pseudo 2-D spectrum when the amplitude of one harmonic is much smaller than others nearby.

where $P_{z_m}(f_1, f_2)$ is the pseudo 2-D spectrum of $z_m(t)$, and the second part is composed of cross terms.

Now, consider the first part of Eq. (8). In polyphonic music, some musical intervals (such as the perfect fifth, perfect fourth, etc.) result in coincidence of harmonics, which is ubiquitous in real-world recordings. Suppose that note m and note n sound simultaneously with such intervals, with fundamental frequencies of $f_{0,m}$ and $f_{0,n}$, partial numbers of H_m and H_n, respectively. In this part, $\{r_i | i = 1, 2\} \in \mathbb{N}$ and $\{l_i, k_i | i = m, n\} \in \mathbb{N}$ are used to indicate the ratio of pitches and pattern locations, respectively. So $f_{0,m}$ and $f_{0,n}$ have the so-called roughly small integer ratio [24], i.e.,

$$f_{0,m}/f_{0,n} \approx r_1/r_2 \tag{9}$$

where r_1 and r_2 are relatively prime, and $\max(r_1, r_2) < \min(H_m, H_n)$.

Thus, we have

$$r_1 f_{0,n} \approx r_2 f_{0,m}. \tag{10}$$

Note m and note n produce their 2-D patterns at $(l_m f_{0,m}, k_m f_{0,m})$ and $(l_n f_{0,n}, k_n f_{0,n})$, respectively, where $1 \leq (l_i, k_i) \leq H_i$, and $i \in \{m, n\}$. When f_1 is multiples of $r_1 f_{0,n}$ and $r_2 f_{0,m}$, i.e.,

$$f_1 = \gamma r_1 f_{0,n} = \gamma r_2 f_{0,m}, \gamma \in \mathbb{N} \tag{11}$$

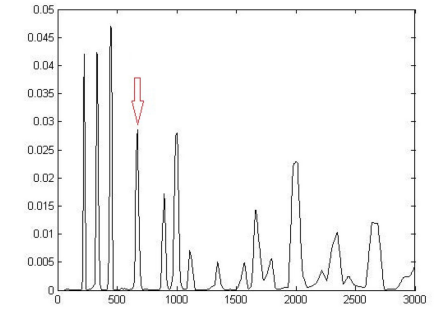
the two patterns both have non-zero values at $(\gamma r_2 f_{0,m}, k_m f_{0,m})$ and $(\gamma r_1 f_{0,n}, k_n f_{0,n})$, respectively. Similar conclusion can be drawn when f_2 is equal to multiples of $r_1 f_{0,n}$ (i.e., $r_2 f_{0,m}$). That's the reason why the overlapping partials from two different notes are mapped to different locations that these two notes belong to. In the pseudo 2-D spectrum plane, even though the harmonic collision is not completely solved, most of the peaks belonging to these two notes are separated apart. For the inharmonic situations, the two patterns do not affect each other.

As an example, the spectrum and pseudo 2-D spectrum of a real-world recording containing A3 (220 Hz) and E4 (329 Hz) are shown in Fig. 3. The third harmonic of A3 and the second harmonic of E4 overlap in the spectrum as shown in Fig. 3(a) (where the red arrow points). But the overlapping partials are located in the two different templates corresponding to these two notes as shown in Fig. 3(b), where the peaks in the red rectangles and green ellipses belong to A3 and E4, respectively. In the Fourier spectrum, one third of the harmonics of A3 overlap with those of E4, and half of the harmonics of E4 overlap with those of A3. However, in the pseudo 2-D spectrum, the spectral peaks of A3 come across those of E4 once for the first 9 peaks and twice for the first 36 peaks in the 2-D template. Correspondingly, the spectral peaks of E4 overlap with those of A3 once for the first 4 peaks and twice for the first 16 peaks in the 2-D template. Hence, the harmonic overlap proportions for both notes decrease greatly. Consequently, the proposed pseudo 2-D spectrum can separate these notes and reduce their interactions.

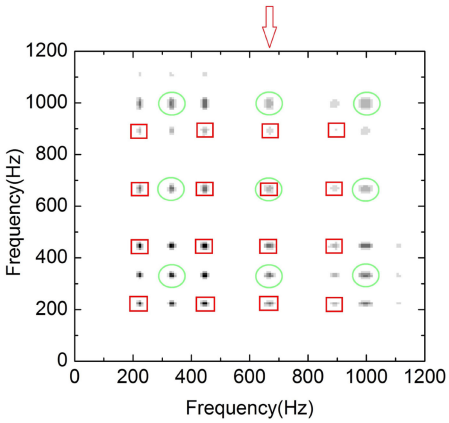
The cross terms in Eq. (8) can be expressed as

$$C_{z_m z_n}(f_1, f_2) = \sum_{l_m=1}^{H_m} a_{m,l_m} \delta(f_1 - l_m f_{0,m}) \cdot \sum_{k_n=1}^{H_n} a_{n,k_n} \delta(f_2 - k_n f_{0,n}). \tag{12}$$

The cross terms are located at $(l_m f_{0,m}, k_n f_{0,n})$, where $m \neq n$, but the 2-D templates corresponding to note m and note n are at $(l_m f_{0,m}, k_m f_{0,m}), 1 \leq (l_m, k_m) \leq H_m$ and $(l_n f_{0,n}, k_n f_{0,n}), 1 \leq (l_n, k_n) \leq H_n$, respectively. Thus, for the inharmonic situations, all cross terms are located out of the 2-D templates. However, for the harmonic situations, where $l_m f_{0,m} = \gamma k_n f_{0,n}$ or $k_n f_{0,n} = \gamma l_m f_{0,m}, \gamma \in \mathbb{N}$, some cross terms belong to either template of note m or note n, and most of the others are located out of the templates, as illustrated in Fig. 3(b). Hence, they do little harm to the 2-D pattern matching.



(a) Spectrum



(b) Pseudo 2-D spectrum

Fig. 3. Spectrum and pseudo 2-D spectrum of an audio containing notes A3 and E4. In (b), the peaks in the red rectangles and green ellipses belong to A3 and E4, respectively. In (a), the third harmonic of the A3 overlaps with the second harmonic of E4, where the red arrow points. Correspondingly, in (b), the components of A3 and E4 are mapped to different locations in the 2-D frequency plane, and the overlapping proportions are reduced greatly for both notes.

The percussions of polyphonic music cover a wide frequency range with no harmonic structure, and they have little influence on the harmonic pattern recognition in the pseudo 2-D spectrum plane if they are not too strong. Hence, the pseudo 2-D spectrum of percussion is not discussed herein.

D. Pseudo 2-D Spectrum Based on Constant-Q Transform

In equal tempered scale, the note pitches are spaced geometrically with one semitone interval. The tolerance for pitch estimation is half semitone (about 3% in frequency) [25]. Fourier transform can be efficiently implemented with constant frequency resolution, but its frequencies don't map musical frequencies efficiently. Hence, constant-Q transform is commonly used in music information retrieval [15], where the frequency are also geometrically spaced.

Given a discrete time signal $x(n)$, its CQT is defined as [26]

$$X_{CQT}(k) = \frac{1}{N_k} \sum_{n=0}^{N_k-1} w(k,n)x(n)\mathrm{e}^{-j2\pi Qn/N_k} \tag{13}$$

where $k = 0, \ldots, K-1$, K is the total number of frequency bins, Q the quality factor, N_k the window length at frequency bin k, and $w(n,k)$ the temporal window function.

The quality factor Q is defined as

$$Q = \frac{f_k}{\Delta f_k} \tag{14}$$

where f_k is the center frequency of the k-th bin, and Δf_k is the bandwidth (or resolution) at f_k.

In CQT, Q is constant, so its bandwidth Δf_k increases with frequency f_k, leading to higher frequency resolutions at lower frequencies and lower frequency resolutions at higher frequencies.

When CQT is adopted for 1-D spectral analysis, the pseudo 2-D spectrum of $x(n)$ is

$$\begin{aligned} P_x(k_1,k_2) &= X_{CQT}(k_1)X^*_{CQT}(k_2) \\ &= \frac{1}{N_{k_1}N_{k_2}} \sum_{n_1=0}^{N_{k_1}-1} w(k_1,n_1)x(n_1)\mathrm{e}^{-j2\pi Qn_1/N_{k_1}} \\ &\quad \cdot \sum_{n_2=0}^{N_{k_2}-1} w^*(k_2,n_2)x^*(n_2)\mathrm{e}^{j2\pi Qn_2/N_{k_2}} \end{aligned} \tag{15}$$

where $P_x(k_1,k_2)$ is the pseudo 2-D spectrum value at the 2-D entry of (k_1,k_2).

Let $\mathbf{P}_x$ and $\mathbf{X}_x$ denote the pseudo 2-D spectrum matrix and CQT vector of $x(n)$, respectively. The Eq. (15) can be rewritten in the matrix form as

$$\mathbf{P}_x = \mathbf{X}_x(\mathbf{X}_x^T)^* \tag{16}$$

where $\mathbf{P}_x \in \mathbb{C}^{K\times K}$, $\mathbf{X}_x \in \mathbb{C}^{K\times 1}$, and $(\cdot)^T$ and $(\cdot)^*$ are the transposition and conjugate operators, respectively.

In musical analysis, the center frequency f_k in Eq. (14) satisfies

$$f_k = f_{low}2^{\frac{k}{b}}, k = 0, \ldots, K-1 \tag{17}$$

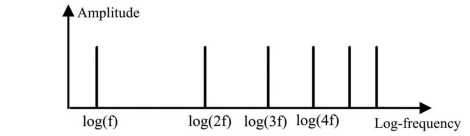
where f_{low} is the lowest frequency, and b is the number of frequency bins per octave.

According to Eq. (17), all harmonic audio signals exhibit the same pattern in the log-frequency scale. For any harmonic signal, the relative positions of all its partials are independent of fundamental frequency, which makes the fundamental frequency estimation can be implemented by shifting the template along the log-frequency axis. The template of one harmonic audio signal with 6 partials in log-frequency scale is illustrated in Fig. 4(a).

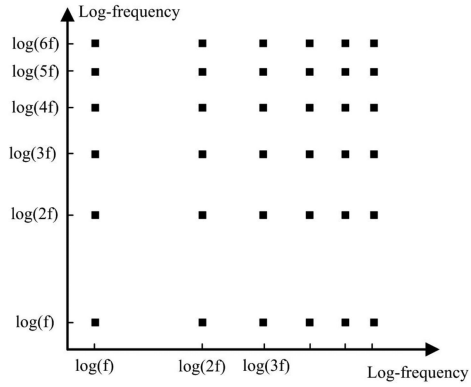
Based on the 1-D template in the log-frequency scale and Eq. (16), the 2-D template of one harmonic audio signal can be built. As an example, the 2-D template with six partials in the 2-D log-frequency plane is illustrated in Fig. 4(b).

III. Pseudo 2-D Spectrum-Based Multi-Pitch Estimation

The definition and analysis of the pseudo 2-D spectrum in Section II reveal that it can be used to solve the multi-pitch estimation problem. In this section, the pseudo 2-D spectrum-based multi-pitch estimation method will be presented, and the block diagram of the proposed method is shown in Fig. 5, where multiple pitches are estimated jointly. First, constant-Q transform is employed to achieve the multi-resolution spectral analysis [27].



(a) 1-D template in log-frequency scale



(b) 2-D template based on the pseudo 2-D spectrum in log-frequency scale

Fig. 4. 1-D and 2-D templates of one harmonic audio signal with six partials in log-frequency scale.

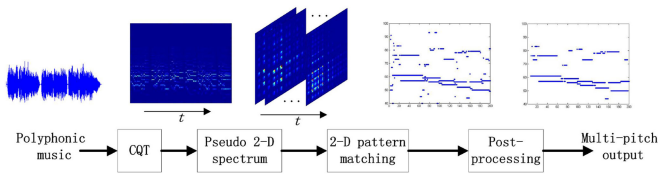


Fig. 5. Block diagram of the pseudo 2-D spectrum-based multi-pitch estimation.

Next, the pseudo 2-D spectrum elaborated in Section II is directly constructed in the frequency domain based on CQT. After that, pattern matching is conducted through cross-correlation. The preliminary pitch candidates are constituted of the spectral peak frequencies whose cross-correlation values are greater than the threshold determined by the maximum cross-correlation value. Finally, pitches of adjacent frames are streamed to form *note contours* - time continuous sequences of F0 estimates that form a note. The isolated pitches are discarded by the post-processing to satisfy the minimum note length limitation, and the remained pitches are refined using the estimates of neighboring frames. The pseudo 2-D spectrum constructed by constant-Q transform, 2-D pattern matching and post-processing will be described in detail in the next subsections.

A. Pseudo 2-D Spectrum Construction

The frame-wise pseudo 2-D spectrum is constructed using CQT, which is a time-frequency representation of an audio signal with varying frequency resolutions for different frequency bands. Given an audio signal $x(n)$, its CQT is first calculated according to Eq. (13). Then, the frame-wise pseudo 2-D spectrum of this signal is built according to Eq. (15).

According to the presentation in Subsection II-D, the pseudo 2-D spectrum constructed by CQT maps sinusoidal components into the 2-D plane, and the pseudo 2-D spectrum is conjugate symmetrical with respect to the quadrant bisector. Hence, the pattern matching can be easily done in the pseudo 2-D plane by shifting the 2-D template along the quadrant bisector, which is exploited for multi-pitch estimation in this work.

B. 2-D Pattern Matching

The harmonic signals exhibit a typical 2-D pattern in the pseudo 2-D spectrum plane, based on which the multi-pitch estimation is explored. There are two conditions that the estimated pitches must satisfy: a) the harmonic structure in the pseudo 2-D spectrum plane matches with the 2-D pattern; b) it has significant peaks in the pseudo 2-D spectrum. According to these two criteria, the 2-D pattern matching is conducted by the cross-correlation, which is calculated as follows. Suppose that there are N_{oct} bins within each octave and H_r harmonics in consideration for each note. Let $\mathbf{Q} = (q_{i,j})$ be a binary matrix with dimension $R_q \times R_q$, where $R_q = 1 + N_{oct}\log_2\lceil H_r \rceil$, and $\lceil \cdot \rceil$ denotes rounding towards positive infinity herein. $\mathbf{Q}$ is a sparse 2-D template matrix in the equal-tempered scale, where $q_{i,j} = 1$, if and only if there are harmonic partials both at the i-th and j-th frequency indices shifting from the fundamental frequency index.

The cross-correlation between the 2-D pattern and the pseudo 2-D spectrum is given by

$$r(k_1, k_2) = \sum_{i=0}^{R_q-1} \sum_{j=0}^{R_q-1} q_{i,j} \, |P_x(k_1 + i, k_2 + j)|. \tag{18}$$

The pseudo 2-D spectrum is conjugate symmetrical with respect to the quadrant bisector, and the maximum of cross-correlation is expected to be located upon the quadrant bisector whose frequency is pitch. Thus Eq. (18) can be rewritten as

$$r(k) = \sum_{i=0}^{R_q-1} \sum_{j=0}^{R_q-1} q_{i,j} \, |P_x(k + i, k + j)|. \tag{19}$$

Theoretically, the partials in polyphonic music result in the spectral peaks at corresponding frequency bins. Hence, there is no need to match the harmonic pattern at each frequency bin. To reduce the computational load, the cross-correlation values expressed by Eq. (19) are only computed at the frequency indices where $|X_{CQT}(k)|$ are spectral peaks. The CQT spectral peak searching is to find the set Ω satisfying $\Omega = \{k \big| |X_{CQT}(k)| > \max(|X_{CQT}(k+1)|, |X_{CQT}(k-1)|)\}$, where $1 < k < N_C$, and the N_C is the maximum index within the frequency range herein.

The pitches whose cross-correlation values are greater than $\lambda r_{\max}$ are taken as preliminary pitches, where $r_{\max}$ is the maximum frame-wise cross-correlation value, and λ is the threshold factor. The frame-wise multiple pitches are preliminarily estimated based on the 2-D pattern matching presented in this subsection.

C. Post-Processing

As the spectrum of polyphonic music is very complex, some spurious pitches are estimated by the above 2-D pattern matching while some pitches are also missed. The post-processing is to determine whether the estimated pitches are valid, and supplement the omitted ones. As Bregman pointed out that western music tends to have notes rarely shorter than 150 ms, and the minimum note length for melody extraction is often set as 100 ms to 150 ms [28], [29]. In this work, the minimum note length is set as 100 ms. The post-processing is conducted following the work [4] with some minor modifications. First, these pitches of adjacent frames belonging to the same note (with frequency difference less than 0.5 semitone) are grouped together to form note contours. Then, the contours whose pitches belong to the same note and gaps shorter than 100 ms are merged. Next, the contours that last shorter than 100 ms are considered as spurious ones and discarded. Finally, the estimated multiple pitches are refined using the estimates of their neighboring frames as [4]. In more detail, the likely errors are first removed according to the pitch histogram within the neighborhood range. After that, the estimates are reconstructed by the original estimates or the weighted average of the original estimates within its neighborhood. The same neighborhood radius is set, i.e., 90 ms.

To summarize, the detailed procedure of the proposed pseudo 2-D spectrum-based multi-pitch estimation method is shown in Algorithm 1.

D. Computational Complexity Analysis

The computational complexity of the proposed pseudo 2-D spectrum-based method is analyzed on a per-frame basis, with CQT dimension B, the CQT bin number per octave N_{oct}, the maximum spectral harmonic number H_r, maximum spectral peak number M and the 2-D template dimension $R_q \times R_q$. It's difficult to evaluate the computational complexity of post-processing, since the number of frame-wise preliminary pitches are unknown and unfixed. Except for post-processing, the computational cost of the proposed method is mainly composed of two parts: the pseudo 2-D spectrum computation and the 2-D pattern matching. The pseudo 2-D spectrum computation costs $O(B^2)$, and the 2-D pattern matching costs $O(MR_q^2)$, where $R_q = 1 + N_{oct}\log_2\lceil H_r\rceil$, and the $\lceil\cdot\rceil$ represents rounding towards positive infinity herein. Hence, the overall computational cost is $C = O(B^2 + MR_q^2) = O(B^2 + M(1 + N_{oct}\log_2\lceil H_r\rceil)^2)$. Generally, $N_{oct}\log_2\lceil H_r\rceil >> 1$, thus the computational cost is approximately equal to $O(B^2 + MN_{oct}^2(\log_2\lceil H_r\rceil)^2)$. Therefore, it can be seen that the proposed method is easy to implement with lower computational complexity.

IV. Experimental Results and Discussions

Some experiments are conducted to evaluate the performance of the proposed method. The evaluation results of the proposed method as well as the typical reference methods are given in this section.

Algorithm 1: Pseudo Two-Dimensional Spectrum-Based Multi-Pitch Estimation.

Input: Polyphonic music samples $x(n), n = 0, \ldots, L-1$.

Output: Multi-pitch estimation results $\hat{\theta}^t = \{F_{0,1}^t, F_{0,2}^t, \ldots, F_{0,N_t}^t\}$, where N_t is the number of pitches at frame t.

(1) Calculate the CQT of polyphonic music according to Eq. (13), with $N_{oct} = 36$, and frequency range from 61.74 Hz to 11025 Hz.

(2) For each frame, do:

① Construct the pseudo 2-D spectrum $P_x(k_1, k_2)$ using Eq. (15).

② Find the set Ω satisfying: $\Omega = \{k \Big| |X_{CQT}(k)| > \max(|X_{CQT}(k+1)|, |X_{CQT}(k-1)|)\}$, where $1 < k < N_C$, and the N_C is the maximum index within the frequency range herein.

③ Compute the 2-D cross-correlation value for each CQT spectral peak at frequency index k using Eq. (19).

④ Rank preliminary pitches according to the cross-correlation values of step ③.

⑤ Reserve the salient pitches whose cross-correlation values are higher than the $\lambda r_{\max}$, where λ is the threshold factor, and $r_{\max}$ is the maximum frame-wise cross-correlation value, i.e., $r_{\max} = \max(r(k))$, $1 \le k \le N_p$, and N_p is the number of pitches in consideration.

(3) Group the adjacent pitches whose frequency distance is less than 0.5 semitone to form note contours.

(4) Merge the note contours whose pitches belong to the same note and note gaps are shorter than 100 ms.

(5) Discard the contours whose lengths are shorter than 100 ms.

(6) Refine the estimated pitches using those of neighboring frames as in [4]. The neighborhood frame radius is 90 ms.

Return: Frame-wise multiple pitches of polyphonic music.

A. Experimental Setup

The parameters in the proposed method are set as follows. (1) A CQT toolbox is used to implement the multi-resolution spectral analysis [27], and the CQT center frequencies are equally distributed in the logarithmic frequency scale between 61.74 Hz and 11.025 kHz with 36 bins per octave. For the pitch estimation tasks, pitches within half semitone range of the ground truth are considered correct. Thus, $N_{oct} = 36$ is enough to satisfy this condition. (2) The possible MIDI pitch range is set between 35 (61.74 Hz) and 96 (2093 Hz), and covers most of the pitches in real-world recordings; (3) The pseudo 2-D spectrum is generated directly based on the CQT, which is logarithmic in frequency scale. The frequency bin interval decreases with the increase of harmonic number. If too many harmonics are taken into account, the partials from other notes may be located in the pattern. Moreover, according to several existing studies, the spectral slopes of music instruments decay 3 dB to 12 dB

per octave [22], [23]. The higher harmonics contribute less to the cross-correlation value. Thus, the harmonic number should not be too big. In the proposed method, H_r is set to be 5, the same as the bispectrum-based method. (4) The 2-D template dimension $R_q = 1 + N_{oct}\log_2\lceil H_r \rceil = 1 + 36 \times 3 = 109$. (5) The maximum time gap for constructing note contours is set to be 100 ms empirically. (6) The minimum note length is 100 ms according to the musical acoustics [28]. (7) The neighborhood frame radius is set to be 90 ms, the same as in [4]. (8) The parameter λ is set using the TRIOS dataset as described in the second paragraph of Subsection IV-D.

B. Evaluation Datasets and Reference Methods

Three datasets are used for evaluation. The first one is Bach10[1], which contains 10 pieces of recordings downloaded from the Internet, lasting from 25 to 41 seconds [4]. There are four parts (Soprano, Alto, Tenor and Bass) of each piece recorded separately with sampling rate 44.1 kHz. The frame-wise average polyphony is 3.56. The second dataset is MAPS-AkPnBcht[2], which comprises 30 pieces of recordings lasting 108 minutes totally [30]. The frame-wise average polyphony is 3.65. The full music pieces are used for evaluation. The frame-wise pitches of both datasets are given in 10 ms intervals. The third dataset is TRIOS[3], used to determine the parameter λ. TRIOS is a score-aligned multi-track dataset [31]. There are five recordings of chamber music trio pieces with their alinged MIDI scores.

The performance of the proposed pseudo 2-D spectrum-based method (P2SB) is compared with eight typical methods, including the constant-Q bispectral analysis (CQBA) [12], spectral peaks and non-peak regions modeling (SPNRM) [4], perceptually motivated (PerM) [9], summing harmonic amplitudes (SHA) [10], adaptive harmonic spectral decomposition (AHSD) [16], the sound space-based spectral factorization (S3F) [14], the note transcription using recurrent neural networks (RNN) [21], and the joint onset and frame estimation using convolutional neural network and bidirectional long short-term memory networks (CBLSTM) [20] methods. These methods are chosen for comparison with the proposed method due to the following reasons. (1) CQBA is also a 2-D spectrum-based method, which makes it more relevant to the proposed method. The other reference methods are all 1-D spectrum-based. (2) PerM and SHA are typical feature-based multi-pitch estimation methods, AHSD and S3F are representative and state-of-the-art spectrogram factorization-based methods, and SPNRM is a typical statistical model-based method. RNN and CBLSTM are two typical deep learning-based methods. (3) The source codes of SPNRM,[4] PerM,[5] SHA,[6] RNN[7] and CBLSTM[8] can be downloaded from the web pages. And the source codes of S3F, CQBA and AHSD can be obtained from the authors.

[1][Online]. Available: http://www2.ece.rochester.edu/projects/air/resource.html

[2][Online]. Available: http://www.tsi.telecom-paristech.fr/aao/en/category/database/

[3][Online]. Available: https://c4dm.eecs.qmul.ac.uk/rdr/handle/123456789/27

[4][Online]. Available: http://www2.ece.rochester.edu/projects/air/publications.html

[5][Online]. Available: https://github.com/tiendung/mutiple-f0-estimation

[6][Online]. Available: https://github.com/tiendung/mutiple-f0-estimation

[7][Online]. Available: https://github.com/CPJKU/madmom

[8][Online]. Available: https://github.com/tensorflow/magenta

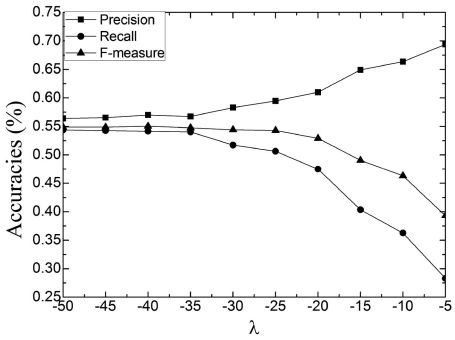


Fig. 6. Performance influence of λ for TRIOS.

C. Evaluation Metrics

The performance of the proposed and eight reference methods is evaluated using MIR_EVAL library [25], [32], [33]. As the main focus of this paper is on the pseudo 2-D spectrum and its application to multi-pitch estimation, the performance evaluation is conducted at frame level. Some evaluation metrics commonly used for multi-pitch estimation are introduced, including precision (P), recall (R), F-measure ($F_{measure}$), substitution error rate (E_{subs}), miss error rate (E_{miss}), false alarm error rate (E_{fa}) and total error rate (E_{tot}). Please refer to [33] for the detailed definitions of the first three metrics and [32] for the others.

D. Evaluation Results and Discussions

In this subsection, the evaluation results and corresponding discussions are provided in detail.

1) Setting of Parameter λ: If a greater λ is taken, the precision will rise at the expense of missing more true positives. And on the contrary, if a smaller λ is taken, the recall rate will go up but precision down. An experiment is conducted to evaluate the influence of λ on the performance using the TRIOS dataset. The experimental results in terms of precision, recall and F-measure are given in Fig. 6. It can be seen that the F-measure decreases with the increase of λ, and the three accuracies all converge to some values around 0.55 with the decrease of this parameter. Moreover, the F-measure varies slightly when λ is smaller than -25 dB. Considering the results on this dataset, $\lambda = -40$ dB is taken for all of the following tests.

2) Superiority of Pseudo 2-D Spectrum Over 1-D Spectrum in Real-World Recordings: To investigate the superiority of the pseudo 2-D spectrum over 1-D spectrum in real-world recordings, the 2-D pattern matching in the proposed method is replaced by the 1-D pattern matching. The parameter λ is also determined using the TRIOS dataset, and set as -45 dB. The 1-D template is also binary. The other procedures and parameters are the same as the 2-D pattern matching, but the pattern matching is done in the 1-D fashion. Then, the modified 1-D spectrum-based method is used for the multi-pitch estimation task on the Bach10 and MAPS datasets.

The F-measures of the 1-D spectrum-based and pseudo 2-D spectrum-based multi-pitch estimation methods are demonstrated in Fig. 7. It can be observed that the pseudo 2-D

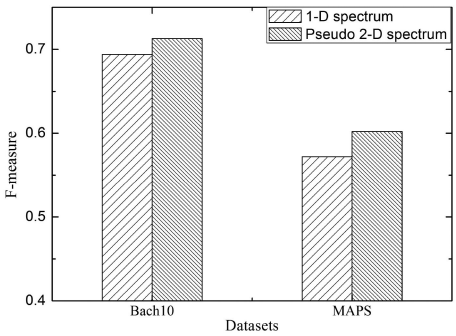


Fig. 7. F-measures of 1-D spectrum-based and pseudo 2-D spectrum-based multi-pitch estimation. The procedures of these two methods are the same except for the signal representation and pattern matching based on 1-D spectrum or pseudo 2-D spectrum.

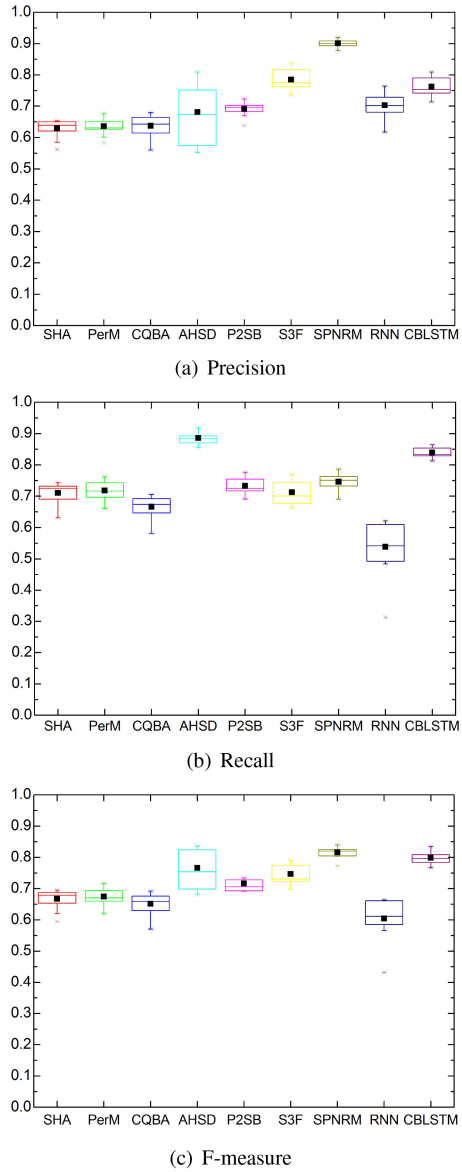
spectrum-based method achieves higher F-measures than the 1-D spectrum-based method on both datasets. Besides, the gap is larger on MAPS whose maximum and average polyphonies are both greater than those of Bach10, respectively.

3) Evaluation Results for Bach10: The precision results of the nine compared methods for Bach10 are shown in Fig. 8(a). It can be seen that SPNRM performs best in terms of precision, which may partly originates from the fact that SPNRM addresses the multi-pitch estimation problem in a statistical framework, where the parameters are learned based on a similar dataset. Among all of the other methods, the proposed method obtains higher average precision than SHA, PerM, CQBA and AHSD, and lower average precision than S3F, RNN and CBLSTM. The box chart of AHSD covers a wider range than the other methods, indicating that its precision varies greater over these excerpts, while the box chart of the proposed method is comparatively oblate, implying its robustness.

The recall results for Bach10 are illustrated in Fig. 8(b). It can be seen that AHSD and CBLSTM achieve higher recall rates than the other methods, while RNN misses a lot of true positives or extracts lots of false positives. The proposed method's performance in terms of this metric is moderate, and the corresponding box chart is still oblate.

To show the integrated results of precision and recall, the F-measure results are reported in Fig. 8(c). It can be seen that SPNRM performs best. The proposed method outperforms SHA, PerM and RNN. Comparing with the bispectrum-based method (CQBA), the proposed method achieves 6.5% higher F-measure rate. RNN and CBLSTM are two deep learning-based methods, and designed for polyphonic piano transcription. In this experiment, their parameters are fine-tuned using TRIOS beforehand. Their results differ a lot, which may due to the fact that CBLSTM has a more complicated architecture (with an acoustic model and a language model), while RNN has a comparatively simpler structure. In addition, the proposed method performs well among the feature-based method, so it might be potentially used in the deep learning framework to further improve its performance.

To offer more insights, the error rates are reported in Table I. From Table I, it can be seen that SPNRM achieves the least total errors for Bach10, while RNN obtains the highest total error rate



(a) Precision

(b) Recall

(c) F-measure

Fig. 8. Performance comparison of different methods for the Bach10 dataset.

TABLE I
ERROR RATES OF MULTI-PITCH ESTIMATION FOR BACH10

Methods	E_{subs}	E_{miss}	E_{fa}	E_{tot}
SHA	0.29	0	0.13	0.42
PerM	0.28	0	0.13	0.41
CQBA	0.32	0.01	0.06	0.39
AHSD	0.08	0.03	0.36	0.47
P2SB	0.12	0.14	0.20	0.46
S3F	0.13	0.02	0.07	0.35
SPNRM	0.07	0.18	0.01	0.27
RNN	0.13	0.35	0.06	0.54
CBLSTM	0.08	0.06	0.15	0.29

which is mostly contributed by miss errors. This observation also affirms that many true positives are missed by this method.

The F-measure values with and without counting octave errors of these compared methods are tabulated in Table II, where the 'woe' is short for 'without counting octave errors'. From

TABLE II
MULTIPLE F0 F-MEASURE WITH (WITHOUT) COUNTING OCTAVE ERRORS ON BACH10

Methods	F-measure	F-measure (woe)
SHA	0.67	0.75
PerM	0.67	0.79
CQBA	0.65	0.76
AHSD	0.77	0.79
P2SB	0.71	0.75
S3F	0.74	0.81
SPNRM	0.82	0.86
RNN	0.60	0.68
CBLSTM	0.80	0.83

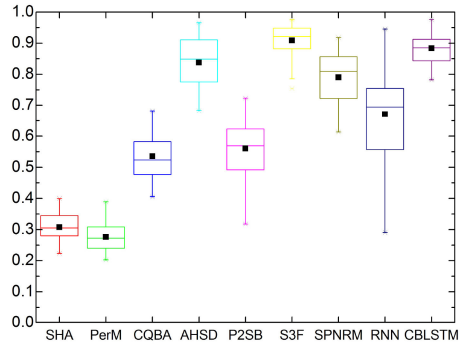
Table II, it can be observed that SPNRM performs best among these methods, and the proposed method achieves lower octave error rate than CQBA.

4) Evaluation Results for MAPS: The performance of these compared methods is also evaluated using MAPS. The precision, recall and F-measure values are shown in Fig. 9. As shown in Fig. 9, the performance of the compared methods varies more diversely for this dataset than Bach10. Additionally, the average F-measure for MAPS is 15.2% lower than that for Bach10. CBLSTM achieves the highest F-measure on this dataset. Again, the proposed method outperforms most 1-D spectrum based method except RNN and CBLSTM, we can infer that combination of the proposed pseudo 2-D spectrum and deep learning techniques might result in better performance, which will be considered in the future work.

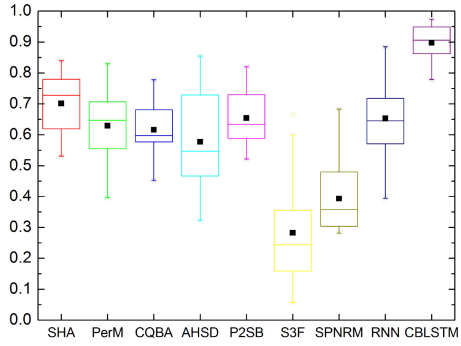
To show the accurate F-measure values and the potential results of these methods, the multiple F0 F-measure results with and without counting octave errors are listed in Table III. It can be seen that the F-measure of CBLSTM is at least 23% higher than the other methods. The two deep learning-based methods (RNN and CBLSTM) perform well, partly because they are trained to transcribe polyphonic piano music, and MAPS is also used to train their parameters. Except for these two methods, the proposed method achieves the second highest F-measure, and it ranks the third if octave errors are ignored. Moreover, CQBA and AHSD also have great potential as seen in this table, surpassing SHA, PerM, S3F and SPNRM. According to the results in Table III, the octave error rate of the proposed method is 3% lower than CQBA.

The corresponding error rates of different methods for MAPS are listed in Table IV. Some total error rates in this table are greater than one since the number of false positives is greater than that of reference pitches. The proposed method outperforms SHA, PerM, CQBA, S3F, SPNRM and RNN, while inferior than AHSD and CBLSTM in terms of total errors. It can be concluded from Fig. 9(c) and Table IV that the proposed method achieves higher F-measure with lower total errors comparing with the bispectrum-based method (CQBA) for MAPS.

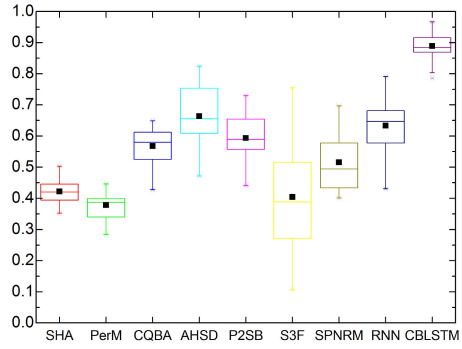
To offer more insights, the ground truth labels and the estimated results of these methods for one excerpt is shown in Fig. 10. This excerpt is taken from 'MAPS_MUS-chpn-p4_AkPnBcht.wav' and its maximum polyphony is 9. It can be seen from this figure that SHA and PerM suffer from many false positives, while S3F and SPNRM miss a lot of true positives.



(a) Precision



(b) Recall



(c) F-measure

Fig. 9. Performance comparison of different methods for the MAPS dataset.

TABLE III
MULTIPLE F0 F-MEASURE WITH (WITHOUT) COUNTING OCTAVE ERRORS ON MAPS

Methods	F-measure	F-measure (woe)
SHA	0.42	0.54
PerM	0.38	0.49
CQBA	0.57	0.67
AHSD	0.66	0.68
P2SB	0.59	0.66
S3F	0.40	0.41
SPNRM	0.52	0.56
RNN	0.63	0.66
CBLSTM	0.89	0.90

CQBA, AHSD and the proposed method perform better than SHA, PerM and S3F methods as a whole for this excerpt. For CQBA, there are some false positives, such as the notes with MIDI number 53 around 3 to 3.5 seconds, as well as some missed pitches, such as the ones with MIDI number 77 between 2.2 and

TABLE IV
ERROR RATES OF MULTI-PITCH ESTIMATION FOR MAPS

Methods	E_{subs}	E_{miss}	E_{fa}	E_{tot}
SHA	0.30	0	1.36	1.66
PerM	0.37	0	1.35	1.72
CQBA	0.20	0.14	0.41	0.75
AHSD	0.04	0.38	0.09	0.51
P2SB	0.17	0.17	0.23	0.57
S3F	0.02	0.70	0.02	0.73
SPNRM	0.08	0.53	0.03	0.64
RNN	0.07	0.27	0.38	0.72
CBLSTM	0.02	0.08	0.10	0.20

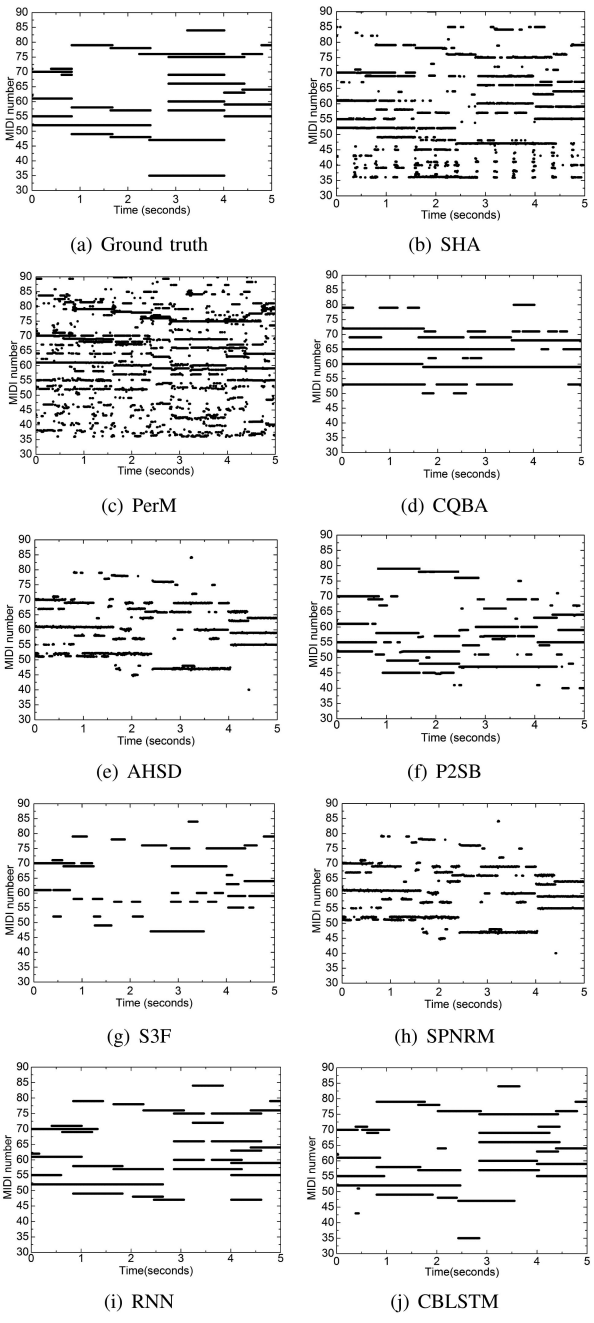


Fig. 10. Ground truth and the estimated multiple pitches by the compared methods for one excerpt. This excerpt is taken from 'MAPS_MUS-chpn-p4_AkPnBcht.wav' and its maximum polyphony is 9.

TABLE V
STATISTICAL SIGNIFICANCE ANALYSIS WITH SOME REFERENCE METHODS IN TERMS OF F-MEASURE

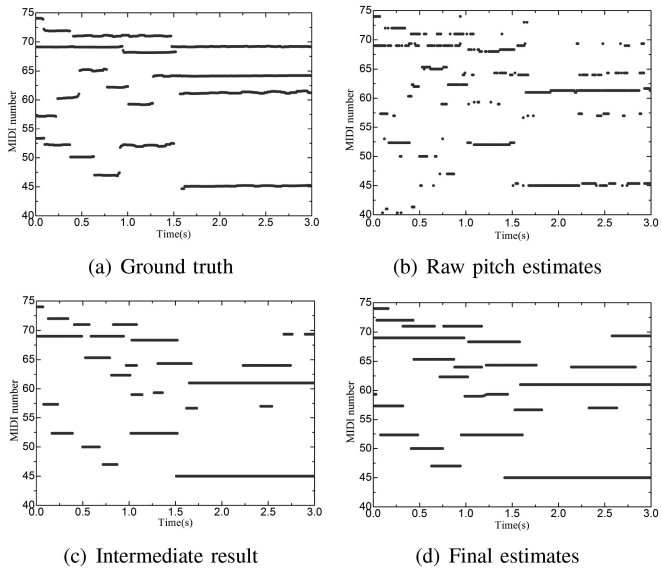
Methods	SHA	PerM	CQBA	S3F	SPNRM
p-value	4.53E-17	6.88E-20	0.14	1.53E-5	2.17E-4

4 seconds. AHSD misses most of the higher frequency pitches. The proposed method also misses the higher frequency pitches around 3 to 4 seconds. RNN and CBLSTM, which transcribe polyphonic piano music at the note level, output less outliers, and their estimation results are less noisy. Almost all of the compared methods miss the note with MIDI number 35 lasting from 2.5 to 4 seconds, but output the note with MIDI number 47 (belonging to the same pitch class but different octaves) for the same interval. Hence, there is still some work to do to distinguish the notes belonging to the same pitch class. Moreover, the observations are in consistent with the accuracies and error rates as depicted in Fig. 9 and Table IV.

As mentioned before, the F-measure integrates both precision and recall, so a paired-sample t-test is performed to compare the F-measure between the proposed method and the others for MAPS. The significance analysis results indicate that AHSD, RNN and CBLSTM perform significantly better than the proposed method (with p-values less than 0.05). Besides, the statistical significance analysis results between the proposed method and the other five methods are tabulated in Table V. It can be concluded that, combining this table with Table III, the proposed method outperforms SHA, PerM, S3F and SPNRM significantly. More specifically, the proposed method outperforms the bispectrum-based method with 18% lower total error rate for MAPS.

5) Effectiveness of Post-Processing: Several post-processing steps are taken in the proposed method, including note merging for gaps shorter than 100 ms, discarding short spurious notes shorter than 100 ms, and pitch refinement using the estimates of neighboring frames. Some experiments are conducted to evaluate their influences on the performance. The experimental results demonstrate that the overall post-processing improves the F-measure 9.2% on Bach10 and 5.7% on MAPS. Among these post-processing steps, pitch refinement using the estimates of neighboring frames contributes the most, note merging contributes slightly less than pitch refinement, while the influence of discarding short notes is marginal.

As an example, Fig. 11 illustrates the effectiveness of post-processing. This excerpt originates from 'BachChorale-4-02-1.wav'. The ground truth is provided in Fig. 11(a). The raw multi-pitch estimation result after 2-D pattern matching is shown in Fig. 11(b). It can be observed that there are some isolated pitches and discontinuous notes in the raw estimates. The remained pitches after note merging and spurious note discarding are demonstrated in Fig. 11(c). It can be seen that the discontinuous notes in Fig. 11(b) have been connected, and the spurious isolated pitches have been removed. The final refined pitches are given in Fig. 11(d). It can be found that post-processing inserts some missed pitches and deletes some spurious ones, indicating the effectiveness of post-processing.



(a) Ground truth (b) Raw pitch estimates

(c) Intermediate result (d) Final estimates

Fig. 11. An example showing the effectiveness of post-processing steps. This example originates from one segment of 'BachChorale-4-02-1.wav'. (a): The ground truth of this segment. (b): The raw pitch estimates after 2-D pattern matching based on the proposed pseudo 2-D spectrum. (c): The intermediate result after note merging and spurious short note discarding. (d): The final multi-pitch estimation result after post-processing.

TABLE VI
RUNTIMES OF DIFFERENT METHODS ON BACH10

Methods	Runtime(second)	Methods	Runtime(second)
SHA	10613.93	S3F	219.45
PerM	6152.05	SPNRM	5152.46
CQBA	491.43	RNN	19.58[9]
AHSD	162.95	CBLSTM	37.74[9]
P2SB	1831.35		

6) Runtime Comparison: The runtimes of these compared methods on Bach10 dataset are tabulated in Table VI. All codes except RNN and CBLSTM run in the MATLAB R2013b on Intel Xeon(R) CPU E5-2640 0 @2.5GHz $\times 18$ with 62.9 GiB memory. The operating system is Ubuntu 14.04 LTS. In this experiment, parallel computing is shut down to compare the overall runtime. For the other experiments, parallel computing is activated for saving time. It can be seen that SHA, PerM and SPNRM cost comparatively more time, while AHSD and S3F need much less time comparing with the other non-deep-learning methods, which may due to their processing at the whole recording level. The proposed method needs more time than CQBA, because the post-processing of CQBA is based on salience smoothing at the recording level, while that of the proposed method is based on note contour construction and refinement dependent on frame-wise pitches. Hence, more efficient strategy for post-processing is still needed to further reduce runtime. As far as the runtime is concerned, the two deep learning methods (RNN and CBLSTM) need few seconds for each recording, since their networks and parameters have been determined during the training stage. The listed runtimes for them are just the test time in Python. If the proposed pseudo 2-D spectrum is incorporated in the deep learning framework, the post-processing time might be saved. This prospective study will be done in our future work. Certainly, the runtimes presented herein roughly reveal the computational complexities, since they also rely heavily on programming tricks, such as looping, matrix operations, and so on.

[9]Runtime for testing on Bach10 using the trained parameters in Python.

V. CONCLUSION

To address the multi-pitch estimation problem, a novel method which utilizes the 2-D pattern of the pseudo 2-D spectrum is presented in this paper. The pseudo 2-D spectrum is first constructed from the constant-Q transform. Next, the 2-D harmonic pattern in the pseudo 2-D spectrum plane is exploited by the cross-correlation between the pseudo 2-D magnitude spectrum and the 2-D template. The salient pitches whose cross-correlation values are greater than the threshold are selected as preliminary pitches. Finally, the preliminary pitches are grouped together to form note contours, and then the contours shorter than the minimum note length limitation are discarded and the remained pitches are refined using the estimates of neighboring frames. The proportions of coincident harmonics in the harmony cases are reduced greatly due to the 2-D mapping, and the proposed method is more robust than the bispectrum-based method for the missing partial situations. Experimental results demonstrate that the proposed multi-pitch estimation method achieves promising performance comparing with the state-of-the-art methods with lower computational cost and a simple post-processing. Compared with the bispectrum-based method, it achieves higher F-measures for both datasets.

Though the proposed method works well on these datasets, there are still several limitations. For example, the average octave error rate is still high, which may originate from the binary matrix used for 2-D pattern matching. Also, pitches are estimated on the per frame basis, thus its runtime is longer than the recording-level estimation methods. Additionally, the post-processing, though not complex, relies on the frame-wise estimated pitches, and the note contour construction and refinement cost much time. The future directions include: (1) replace the binary 2-D matrix with a weighted 2-D matrix for the 2-D pattern matching, similar to the weighted summation in the 1-D spectrum to reduce octave errors; (2) pursue more sophisticated streaming strategies to improve its ability to estimate concurrent pitches belonging to the same source, and to extract the pitches with weaker intensities; (3) incorporate the pseudo 2-D spectrum into other frameworks, such as deep learning, statistical methodologies or nonnegative tensor factorization.

APPENDIX
PROPERTIES OF THE PSEUDO 2-D SPECTRUM

There are four properties of the pseudo 2-D spectrum, which are provided as follows.

1) Conjugate Symmetry:

$$P_x^*(f_1, f_2) = X^*(f_1)X(f_2) = P_x(f_2, f_1). \tag{20}$$

Therefore, the quadrant bisector of the pseudo 2-D spectrum is the power spectrum of the signal $x(t)$, and the pseudo 2-D spectrum is conjugate symmetrical with respect to the quadrant bisector.

2) Time Shifting: The pseudo 2-D spectrum of time-shifted signal $x(t \pm t_0)$ is

$$\begin{aligned} &P[x(t \pm t_0)] \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x(t \pm t_0) x^*(\tau \pm t_0) e^{-j2\pi(f_1 t - f_2 \tau)} \mathrm{d}t \mathrm{d}\tau \\ &= X(f_1) e^{\pm j2\pi f_1 t_0} X^*(f_2) e^{\mp j2\pi f_2 t_0}. \end{aligned} \tag{21}$$

Hence, the following time shifting property holds

$$P[x(t \pm t_0)] = P_x(f_1, f_2) e^{\pm j2\pi(f_1 - f_2)t_0}. \tag{22}$$

According to this property, the pseudo 2-D spectrum of time-shifted signal has the same amplitude spectrum but different phase spectrum with the original one. Hence, for the short-time stationary signal like polyphonic music, the start time of each frame makes no difference to the amplitude of pseudo 2-D spectrum.

3) Frequency Shifting:

$$\begin{aligned} &P[x(t) e^{\pm j2\pi f_0 t}] \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x(t) e^{\pm j2\pi f_0 t} x^*(\tau) e^{\mp j2\pi f_0 \tau} e^{-j2\pi(f_1 t - f_2 \tau)} \mathrm{d}t \mathrm{d}\tau \\ &= X(f_1 \mp f_0) X^*(f_2 \mp f_0). \end{aligned} \tag{23}$$

So we have

$$P[x(t) e^{\pm j2\pi f_0 t}] = P_x(f_1 \mp f_0), f_2 \mp f_0). \tag{24}$$

Due to this property, frequency modulation results in a horizontal and vertical shift of a two-dimensional template in the pseudo 2-D spectrum plane.

4) Marginal Property:

$$\int_{-\infty}^{+\infty} P_x(f_1, f_2) df_1 = x(0) X^*(f_2) \tag{25}$$

$$\int_{-\infty}^{+\infty} P_x(f_1, f_2) df_2 = x^*(0) X(f_1) \tag{26}$$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P_x(f_1, f_2) df_1 df_2 = |x(0)|^2. \tag{27}$$

According to the marginal property, the time domain signal $x(t)$ can be synthesized in terms of $P_x(f_1, f_2)$ through either of the following two equations

$$x(t) = \frac{1}{x^*(0)} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P_x(f_1, f_2) e^{j2\pi f_1 t} df_1 df_2 \tag{28}$$

$$x(t) = \frac{1}{x^*(0)} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P_x^*(f_1, f_2) e^{j2\pi f_2 t} df_1 df_2. \tag{29}$$

The $x^*(0)$ in equations (28) and (29) is constant, and can be removed if $x(t)$ is a real signal.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and the editor for his review of the manuscript. The authors also would like to thank the authors of the reference methods for providing their codes.

REFERENCES

[1] M. W. Hansen, J. R. Jensen, and M. G. Christensen, "Estimation of fundamental frequencies in stereophonic music mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 296–310, Feb. 2019.

[2] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 11, no. 6, pp. 804–816, Nov. 2003.

[3] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: Challenges and future directions," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 407–434, 2013.

[4] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2121–2133, Nov. 2010.

[5] K. Shibata *et al.*, "Joint transcription of lead, bass, and rhythm guitars based on a factorial hidden semi-Markov model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 236–240.

[6] J. A. Moorer, "On the transcription of musical sound by computer," *Comput. Music J.*, vol. 1, no. 4, pp. 32–38, 1977.

[7] M. Akbari and H. Cheng, "Real-time piano music transcription based on computer vision," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2113–2121, Dec. 2015.

[8] C. T. Lee, Y. H. Yang, and H. H. Chen, "Multipitch estimation of piano music by exemplar-based sparse representation," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 608–618, Jun. 2012.

[9] A. P. Klapuri, "A perceptually motivated multiple-f0 estimation method," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2005, pp. 291–294.

[10] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2006, pp. 216–221.

[11] L. Su and Y. H. Yang, "Combining spectral and temporal representations for multipitch estimation of polyphonic music," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1600–1612, Oct. 2015.

[12] F. Argenti, P. Nesi, and G. Pantaleo, "Automatic transcription of polyphonic music based on the constant-Q bispectral analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1610–1630, Aug. 2011.

[13] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[14] E. Benetos and T. Weyde, "Multiple-f0 estimation and note tracking for MIREX 2015 using a sound state-based spectrogram factorization model," *Music Inf. Retrieval Eval. Exchange*, vol. 1, no. 1, pp. 1–3, 2015.

[15] E. Benetos, S. Cherla, and T. Weyde, "An efficient shift-invariant model for polyphonic music transcription," in *Proc. Int. Workshop Mach. Learn. Music*, 2013, pp. 1–4.

[16] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[17] V. Arora and L. Behera, "Multiple F0 estimation and source clustering of polyphonic music audio using PLCA and HMRFs," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 2, pp. 278–287, Feb. 2015.

[18] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 5, pp. 927–939, May 2016.

[19] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2016, pp. 475–481.

[20] C. Hawthorne *et al.*, "Onsets and frames: Dual-objective piano transcription," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2018, pp. 50–57.

[21] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 121–124.

[22] K. Dressler, "Pitch estimation by the pair-wise evaluation of spectral peaks," in *Proc. AES 42nd Int. Conf.*, Jul. 2011, pp. 1–10.

[23] C. D. Tsang and L. J. Trainor, "Spectral slope discrimination in infancy: Sensitivity to socially important timbres," *Infant Behav. Develop.*, vol. 25, no. 2, pp. 183–194, 2002.
[24] J. Sundberg, "Perception of singing," *Psychol. Music*, vol. 1, no. 1, pp. 171–214, 1999.
[25] C. Raffel, B. Mcfee, and E. E. J. Humphrey, "MIR_EVAL: A transparent implementation of common MIR metrics," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2014, pp. 367–372.
[26] Brown and C. Judith, "Calculation of a constant Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, 1991.
[27] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A MATLAB toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Proc. AES 53rd Int. Conf. Semantic Audio*, Jan. 2014, pp. 1–8.
[28] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: MIT Press, 1999.
[29] R. Paiva, T. Mendes, and A. Cardoso, "Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness," *Comput. Music J.*, vol. 30, no. 4, pp. 80–98, 2006.
[30] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, 2010.
[31] J. Fritsch, "High quality musical audio source separation," Master's Thesis, Dept. Center Digit. Music., Queen Mary Univ. London, London, U.K., 2012.
[32] G. E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP J. Adv. Signal Process.*, vol. 8, no. 1, pp. 154–154, 2007.
[33] "Music Information Retrieval Evaluation eXchange (MIREX)," 2007. [Online]. Available: http://www.music-ir.org/mirex/wiki/2007:Multiple_Fundamental_Frequency_Estimation_%26_Tracking_Results

Weiwei Zhang received the Ph.D. degree in signal and information processing from the School of Information and Communication Engineering, Dalian University of Technology (DUT), Dalian, China, in 2019. She is currently a Lecturer with the Information Science and Technology College, Dalian Maritime University. Her research interests include music information retrieval and speech processing.

Zhe Chen (Senior Member, IEEE) received the B.S. degree in electronic engineering, the M.S. and Ph.D degrees in signal and information processing from the Dalian University of Technology (DUT), Dalian, China, in 1996, 1999 and 2003, respectively. He joined the Department of Electronic Engineering, DUT, as a Lecturer in 2002, and became an Associate Professor in 2006. His research interests include speech processing, image processing and wideband wireless communication.

Fuliang Yin was born in Fushun city, Liaoning province, China, in 1962. He received the B.S. degree in electronic engineering and the M.S. degree in communications and electronic systems from the Dalian University of Technology (DUT), Dalian, China, in 1984 and 1987, respectively. He joined the Department of Electronic Engineering, DUT, as a Lecturer in 1987 and became an Associate Professor in 1991. He has been a Professor at DUT since 1994, and the Dean of the School of Electronic and Information Engineering, DUT from 2000 to 2009. His research interests include speech processing, image processing and broadband wireless communication.