PipaSet and TEAS: A Multimodal Dataset and Annotation Platform for Automatic Music Transcription and Expressive Analysis dedicated to Chinese Traditional Plucked String Instrument Pipa

Yuancheng Wang, Yuyang Jing, Wei Wei, Dorian Cazau, Olivier Adam, Qiao Wang

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 00.0000/ACCESS.2021.DOI

PipaSet and TEAS: A Multimodal Dataset and Annotation Platform for Automatic Music Transcription and Expressive Analysis dedicated to Chinese Traditional Plucked String Instrument Pipa

YUANCHENG WANG[1], YUYANG JING[2], WEI WEI[3], DORIAN CAZAU[4], OLIVIER ADAM[5], QIAO WANG[1]

[1]School of Information Science and Engineering, Southeast University, China (e-mail: {wangyuancheng, qiaowang}@seu.edu.cn)
[2]Nanjing University of the Art (e-mail:jingyuyang@nua.edu.cn)
[3]Conservatory of Music, XiaoZhuang University, Nanjing, China.(e-mail: weiwei@njxzc.edu.cn)
[4]Institute of Mines-Télécom Atlantique, Lab-STICC, UMR 6285, Brest, CNRS, France (e-mail: dorian.cazau@imt-atlantique.fr)
[5]Institute of Jean Le Rond d'Alembert, Sorbonne Univerisity, UMR7190, CNRS, Paris, France (e-mail: olivier.adam@sorbonne-universite.fr)

Corresponding author: Yuancheng Wang (e-mail: yuanchengwang@seu.edu.cn).

ABSTRACT Music Information Retrieval (MIR) develop rapidly these years, Automatic Music Transcription (AMT) and Expressive Analysis is increasingly gaining momentum on both Western and non-eurogenic music. However, the annotated datasets for non-eurogenic instruments remain scarce on quantity and feature diversity so that general evaluations and data-driven models on various tasks cannot be well-explored. As one of the most popular traditional plucked string instruments in Asia barely analyzed in MIR community, pipa has lots of distinctive national and local characteristics mainly including 4 classes of sophisticated playing techniques greatly enhancing the music expressiveness. Our work aims to systematically clarify an efficient procedure for the multi-modal pipa dataset creation which consists of audio, musical notations and multi-view videos of Chinese traditional solos. The use of 4-track string vibration signals captured by optical sensors paves a path to high quality annotation. A Transcription and Expressiveness Annotation System (TEAS) is transparently implemented to ensure the scalability of dataset. Finally, a series of existent and new MIR tasks enabled by this dataset are enumerated to explore in the future.

INDEX TERMS Multimodal Dataset Creation, Automatic Music Transcription, Playing Technique Analysis, Optical Sensoring, Annotation System.

I. INTRODUCTION

A. BACKGROUND AND TREND TO THE MULTIMODAL AUTOMATIC MUSIC TRANSCRIPTION AND EXPRESSION ANALYSIS

Automatic Music Transcription (AMT) and polyphonic music analysis have emerged for almost 50 years [1]. As a fundamental task of the MIR, it's still very challenging on the basis of results in the MIR community [2]. The purpose of AMT system narrowly defined in [3] aims to indicate basically a low-level quadruple i.e. pitch, intensity, onset time, and duration of each sound that was played. [4] summarizes the frame-level, note-level, stream-level and notation-level transcriptions in which the notation-level one, also referred to as Complete Notation Transcription (CNT), converts an audio signal into some form of music notations, such as sheet music [5]. However, the music notations cannot completely reflect the characteristics of different performers.

The prior knowledge, such as the music, acoustic features of particular instruments, is always taken into account at the beginning of the research. The AMT studies on piano [6] and guitar [7] introduce the prior knowledge of the occidental instruments and music theory, the others on vocals in Indian

music [8], Ney and Tanbur in Turkish Makam music [9] and Marovany zither in Madagascar music [10] investigate the the non-eurogenetic music transcriptions which manifest the characteristics distinct from the assumptions of the occidental music conventions, e.g. non-equal temperament, free tempo, specific playing techniques and musical notations which are highly relative to the regional culture, styles and genres. These works greatly expand the scope of the MIR research and have developed an inter-disciplinary field, computational ethnomusicology [11], [12]. De-centering the western music has become a path to the cultural diversity in MIR field [13].

In recent decade years, the expression analysis [14] and the multimodality [15] have become promising trends MIR community. Depicting a rich variety of individual emotions, styles and culture, the expression analysis serves to playing technique detection and parameter estimation in which the latter promotes the concept of CNT. The multi-modalities provide a wider perspective and more direct perception to the music features, particularly in the complex scenarios [16], [17] (for more details, please see section IV.C). Audio-only playing technique analyses on electric guitar [18]–[23], bowed strings [24]–[26] and Chinese Bamboo flute [27], [28] enrich the functionality in a single angle. The multimodality analysis covers the synchronization, similarity, time-dependent representation and classification in MIR tasks [29], [30], most of works usually focus on audio, video, score modalities which can accompany with expression analysis [31], [32]. More generally, the sensoring components like the key pressure transducer in piano [33] and the strain gauge to acquire bowing motion for string quartet [34] seen as the physical layer modalities are commonly applied to efficient music analysis and annotation.

B. DESCRIPTION OF PIPA AND PIPA MUSIC

Pipa, the undisputed head of the traditional plucked string instruments in China, has over 2000 years of history. The most ancient straight-necked pipa can be traced back to the Qin Dynasty. The bent-neck pipa was then re-introduced through the Silk Road from Persia and horizontally played as the on-the-horse musical instrument. With the evolution of the playing techniques and repertories, the pentatonic pipa became obliquely played with plectrum then vertically played with fingers at Tang Dynasty. It has been progressively integrated into Chinese music cultures and spread to the other countries in East and Southeast Asia as the most popular plucked string instrument.

1) Pipa Construction

In this paper, we are interested in the modern Chinese pipa standardized in the middle of last century which are chromatic and unified as pear shape. The fake nails substitute for fingernails and plectrums so as to more agilely perform the crisp sound. The research object in our work is a six-ledged (Xiang, 相), twenty-four fretted (Pin, 品), four-stringed pipa

FIGURE 1: The six ledged, twenty-four fretted, four-stringed Pipa (Left) and celluloid-made fake nails (Right) used in our work.

String index	Lowest note	Highest note	Number of frets/ Playable frets	Whole tone fret
str1	A3	E6	31/31	Yes
str2	E3	B5	31/31	Yes
str3	D3	D5	31/25	No
str4	A2	#F4	25/22	No

TABLE 1: The pitch ranges for each string under Pipa standard tuning method.

[35] and a suit of celluloid-made translucent fake nails (See Figure 1). Ledges count into frets in what follows.

Similarly to the acoustic guitar, the modern pipa is often equipped by the steel strings with different thickness in which the 3 thick strings are copper string-wrapped or nylon-wound. The thickness and materials physically determinate the amplitude distribution and inharmonicity of harmonic series [36] which allow the estimation of the string, fret and plucking positions from audio [37], [38]. Furthermore, according to the pipa construction and performance rules in [39], the playable pitch range are shown in Table 1. Concretely, the typical tuning method, Standard D, i.e. A2, E3, D3, A3, is chosen upon the open strings to simplify the following recordings. Notice that the pitch interval between two adjacent frets corresponds to approximately a semitone except those of the lowest 2 frets of 1st and 2nd strings, the register ranges thus from A2 to E6. Considering both register and partials of high tones, the frequency range covering at least from 110 Hz to 14000 Hz is recommended to take up the subsequent research.

FIGURE 2: Pipa notation of Swan excerpt.

2) Music Notations and Taxonomy of Playing Techniques

The ancient pipa notations, e.g. GongChe (工尺) notations, contains the symbols written with Chinese characters and radicals, traditional musicians pay attentions to oral transmission which more or less deprives of key music features in ancient notations like tempo and rhythm. The modern pipa notation is a tonic sol-fa notation combining with both Western and Chinese musical symbols of which a font example could be found in [40]. Besides the basic normal plucks, the fingerings, string number, playing techniques are versatile and especially highlighted in pipa music. Shown in Figure 2, lots of articulation symbols all over the notation reflect the indispensable needs for pipa music transcription. The nomenclature of pipa, originated from two fingerings Pi and Pa, respectively meaning the plucking forwards and backwards, implies the diversity and complexity of playing techniques.

Above all, over 60 techniques [41] must be categorized into the identical or similar notions in the sense of Western music theory. Firstly, we distinguish the basic fingerings and articulations in pipa music. The plucking string, position, finger, direction, inclination for right hand and the termination position (fingertips upon the fret or beside the fret) for left hand are classified into the basic fingerings, so are the techniques constituted by basic pluck(s) like double-pluck (ShuangTan, 双弹) which represents the simultaneous plucks on two different strings. Vibrato, sliding, tremolo, strumming and those perfectly correspond to the articulation types in Western music framework. Furthermore, since the performance parameters have a major impact on a listener's perception of the music [42], the names are often accompanied by parameters in pipa techniques, e.g. the triple tremolo (SanLun, 三轮) which contains 3 intensive plucks sequentially played with index, middle and annular fingers. This signifies not only the detection to indicate the articulation intervals but also the parameter estimation within the intervals must be involved in pipa analysis framework. The left hand techniques (i.e. vibrato, trilling, bending, sliding, harmmer-on/pull-off), the right hand ones (i.e. tremolo, strumming), as well as their corresponding parameters are gathered in Table 2 (See Section III for more details). Harmonics, staccato, different percussive and noise techniques like string scratch, slapping and board flick, twisting (Jiao, 绞) and combining (Bing, 并) count into the 'Others' item in the table.

Categories	Articulations	Feature parameter(s)
Pitch fluctuation	Vibrato, trilling, bending	Extent, rate.
Pitch transition	Sliding, hammer-on, pull-off, pull, push	Inflection point, growth rate.
Tremolo	Wheel, rolling, shaking	Plucking number, rate, dynamic.
Strum	Strumming, arpeggio	Rate, strings of departure and destination.
Others	Harmonics, percussion, special noises	–

TABLE 2: Articulation information in dataset.

C. PRINCIPLE OF DATASET CREATION AND ORGANIZATION OF THE PAPER

The collection and analysis framework for PipaSet follows the principles described by Su and Yang [43] and Xi [32]:

- **Generality:** This dataset covers realistic, complex, polyphonic musical phrases, the performers' motions and scores without any abridgement on original pipa repertoires. The self-interpretation on tempo, rubato, dynamic variations and ornamentation is allowed to highlight the natural charm of Chinese music. The recording progress is developed in popular style.
- **Quality:** In order to preserve nuances in the performance, including the pitch deviation and articulations, we elaborate a vibration sampling device for individual string and provide multiple annotation formats to ensure high quality annotations.
- **Efficiency:** The annotation and visualization are so automated through the highly interpretable mapping algorithms that the music experts could focus on manual corrections in most of cases. PipaSet can be easily extended for this reason.
- **Cost:** The equipment is either common for recording studio or very affordable to acquire like a small piece of Printed Circuit Board (PCB) with optical switches.

The remainder of paper is organized as follows. The multimodal recordings, including the optical sensors to capture vibration signal from individual string, are present in Section II. Section III shows an annotation platform implemented using MATLAB GUI that benefits from optical vibration recordings. In Section IV, dataset statistics, detailed acoustic and music features for target pipa and repertories used in our study and applications empowered by this dataset are present. Finally, Section V contains the conclusion and the future work for this dataset.

II. MULTIMODAL RECORDINGS

The dataset creation, always motivated by the research objectives, has begun to extend to many non-audio modalities in MIR tasks. The recording devices and music instruments for

Datasets	Microphone(s)	Camera(s)	Sensor type(s)	Instrument/Genre
MAPS [33]	Single-channel		Key Pressure Transducer	Piano
MAESTRO v3 [112]	Dual-channel		Key/pedal Pressure Transducer	Piano
Yang [25]	Single-channel			Erhu, Violin.
EEP [34]	Single-channel contact		Piezoelectric pickup/Strain gauge/ Electromagnetical position sensor	String quartet
CBF dataset [27], [28]	Single-channel			Chinese bamboo flute
Multimodal Guitar [31]	Dual-channel contact	Multi-view		Electric guitar
Guitar playing technique [20]	Dual-channel			Electric guitar
IDMT-SMT [21]	Dual-channel			Electric guitar
URMP [32]	Single-channel	Single-view		Multi-instruments
Marovany Zither [46]	Single-channel		Optical switch	Marovany zithers
GuitarSet [45]	Single-channel		Magnetic pickup	Guitars
CTIS dataset [76]	Dual-channel	Signle-view		Multiple Chinese instruments
PipaSet	Dual-channel + Camera embedded microphones	Multi-view	Optical Switch	Pipa

TABLE 3: Recording devices and music instruments of datasets discussed in this paper.

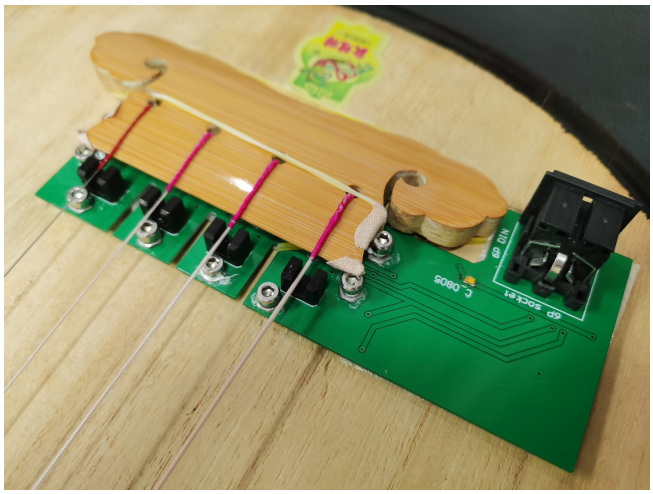


FIGURE 3: The optical sensors used to capture string vibrations in our work.

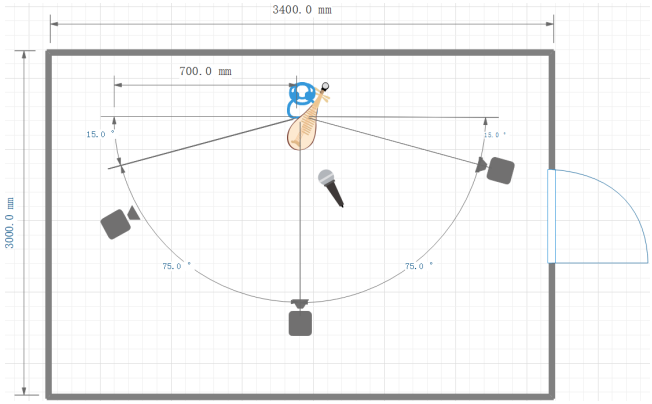


FIGURE 4: Player and audio-video equipment layout in anechoic room.

some related datasets can be found in Table 3. The manual annotation from polyphonic and complex audio recordings has low-efficiency and uncontrollable deviation for different annotators, the convenience and resolution for high-quality annotation must be considered within selected modalities. In this section, the sensors to capture the signal from individual string, the device deployment for pipa polyphonic recording are described as follows.

1) Single-string Vibration Recording using Optical Sensors

The string vibration sensing, physically filtering the external noise like those from fake nails and knocks, can be mainly divided into three types. Usually located on the bridge, the piezoelectric pickups [34], [44] for guitar and violin are not adapted to the pipa construction in which the strings are directly dragged through the tailpiece without need of saddle support. Xi [45] utilizes the hexaphonic pickups to capture the vibration but no similar products for our instrument. Cazau [46] provides a simple reference optical system for unmovable Marovany zithers using an external support and proves the distortion impact of optical sensors is minor and acceptable. To achieve detachability, we customize a system using small optical switches mounted within the small gaps between the strings and the soundboard unlike those on Zithers so as to prevent from disturbing the performance (See Figure 3).

The principle of the optical switches is to detect the surface shadowed by strings, the string vibrate within a small sensing interval to avoid the clipping effect and waveform warping. The switches are thus placed close to tailpiece to ensure the absence of low frequency wave nodes and the amplitude range appropriate to the optical switch. A lobulated PCB is designed to independently adjust switch heights to strings using the fine-tuning screws. The waterproof tape wrapping around the screws avoids the loose noise.

2) Multimodal Device Deployment and Synchronization

A pair of sE8 condenser microphones [47] with a A/B stereo mode is placed at 0.3 m of distance pointed to the sound hole under the tailpiece. The 6-track analog signals from optical switches and microphones are all connected to a single RME fireface UFX sound card [48] to ensure the synchronization. All 6-channels signals are sampled with rate of 48 kHz and depth of 24. In addition, 3 industrial cameras using Sony IMX214 CMOS image chips are selected to capture

the multi-view finger behaviors to avoid the occlusion from both hands with the sampling rate of 30 fps and resolution of 1080p. Two adjacent ones have 75 degree difference oriented to the performer with 0.7 m of distance. Landtop Faist anechoic chamber provides an ambience preventing from the spatial acoustics during recording, in which the green background curtain facilitates video processing like character matting. The layout is shown in Figure 4. The different colors of nail stickers adhered on the fake nails of each finger are chosen as visually conspicuous markers (See right plot of Figure 1). Synchronization among different devices is still challenging for multi-modal recordings. The externally triggered cameras are often misaligned with the audio system. In our study, the weak microphones embedded in cameras are activated to align audio-visual devices using "synchronize clips" function of the Final Cut Pro software.

III. TEAS: A TRANSCRIPTION AND EXPRESSIVENESS ANNOTATION SYSTEM

The annotation is a time-consuming and labor-intensive work which often requires the field expertise. In order to maximize the efficiency and automation, we greatly enhance the AVA platform [25] which only focuses on the continuous pitch, vibrato and portamento. The design notion of TEAS platform is driven by principles below:

- The pipeline is clear, concise and reserves the vast majority of performance-level rather than note-level attributes.
- Distinct from transcription system, the annotation system aims to a new instrument and accompanied with high interpretability and controllability (low-parameter). A suitable low-precision for mapping algorithms is favorable as the deletion is easier than adding.
- The user interaction modules, e.g. visualization, efficient manual correction strategies and import/export in different formats, need to be involved at arbitrary step.

Shown in Figure 5, the annotation usually operates from the the low-level to the high-level music representations as a progressive dimension reduction procedure. First we try to remove the inevitable inter-stringed resonance with Kernel Additive Method Interference Reduction (KAMIR) source separation method [49] referred to [45]. Embedded in TEAS, KAMIR module decomposes 4 channels of string vibration signals into 4 tracks of debleeded signals, meanwhile the diagonal elements in initial interference matrix are configured as 1 and 0.8 for the rest to ensure channel matching. Although the MMSE-LSA adaptive filter [50] could also reduce the parasite noise from electronics and the mutual resonances of strings [46], the separated signal sounds much more likely to the realistic audio recording. Figure 6 shows an example of the raw optical signal and debleeded signal of first string, the significant resonance note removal from another strings can be found located between 6.8-7.8s.

Similarly to Tony [51] and Melodyne [52] softwares, the monophonic AMT modules in TEAS indicate boundaries, pitch contours, note segments and corresponding intensity

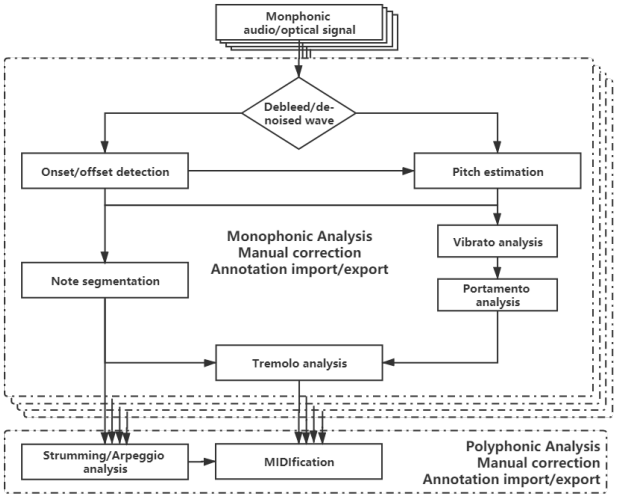


FIGURE 5: The flowchart of TEAS used to assist the transcription and expressive annotation.

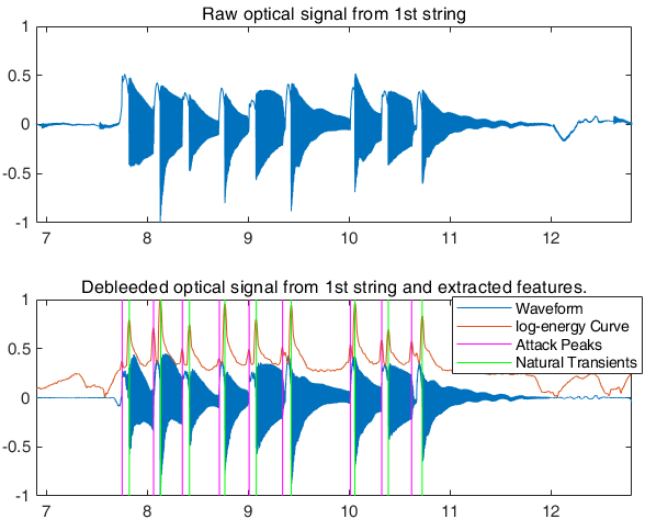


FIGURE 6: Raw optical signal and debleeded signal of 1st string for **Jasmine Flower** excerpt.

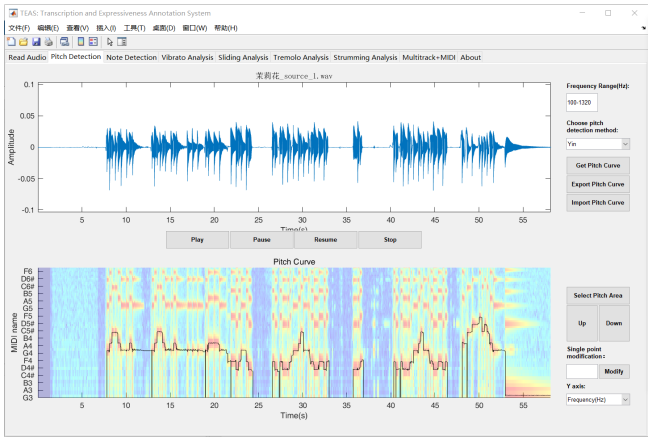
on each string as the second step. Furthermore, playing technique analysis modules using the above-achieved features are elaborated. The screenshots of TEAS are shown in Figure 7.

A. MONOPHONIC AUTOMATIC MUSIC TRANSCRIPTION

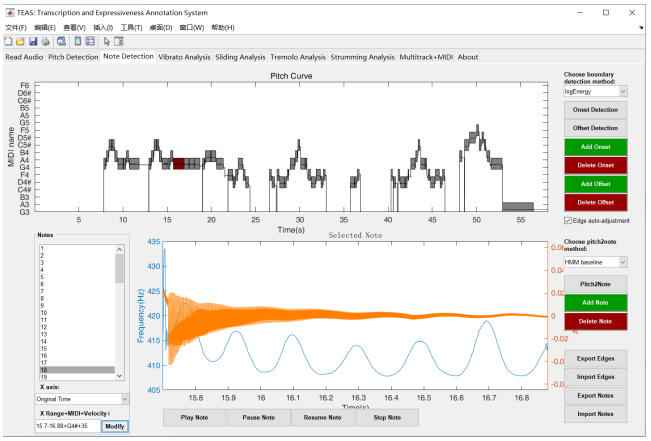
1) Pitch Estimation and Tracking

Pitch detection always confronts with the challenges of the noise and time-variance in music and speech recordings. The traditional pitch tracking algorithms can be mainly classed into non-parametric and parametric approaches. In most of recent dataset creation works, the pitch contours are extracted by the non-parametric state-of-the-art pYIN [53].

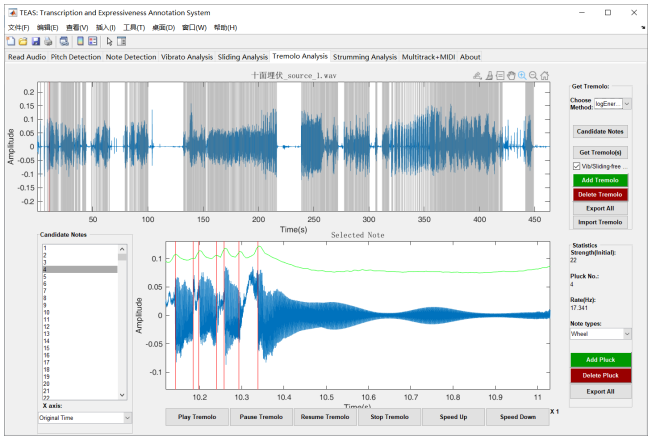
A parametric method, Bayesian pitch tracker using Nonlinear Least Square (BNLS) [54], is provided. The input signal is sliced into N frames and expressed as $\mathbf{Y}_N =$



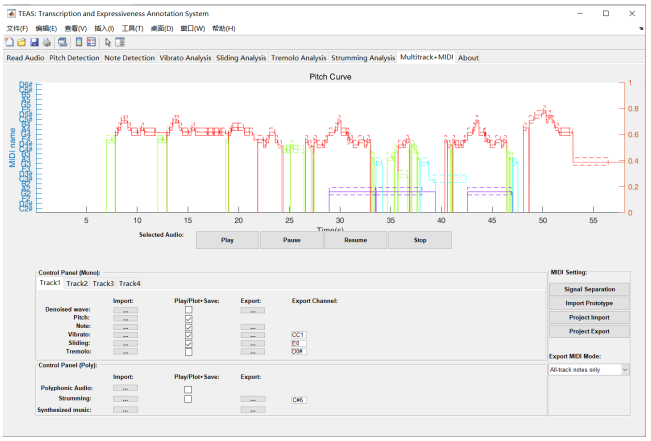
(a) Pitch detection for debleeded first string vibration signal of **Jasmine Flower** recordings



(b) Note segmentation for debleeded first string vibration signal of **Jasmine Flower** recordings



(c) Tremolo analysis for debleeded first string vibration signal of **Ambush from Ten Sides** recordings



(d) Annotations overview for all tracks of **Jasmine Flower**

FIGURE 7: TEAS screenshots for 2 pipa masterpieces **Ambush from Ten Sides** and **Jasmine Flower**.

$[\mathbf{y}_1, ..., \mathbf{y}_N]$ where the n^{th} frame $\mathbf{y}_n = [y_{n,1}, ..., y_{n,M}]^T$ has M samples and the $y_{n,m}$ is zero-centered in n^{th} frame. Assuming the partial frequencies as the integral multiples of the fundamental frequency, the main idea of parametric methods is to reconstruct the $\mathbf{y}_n$ with an ensemble of Fourier bases in terms of harmonic series within preset register which can be formulated as follows:

$$\begin{aligned}\mathbf{y}_n &= u_n \sum_{k=1}^{K} Z(\omega_{k,n}) a_{k,n} + \varepsilon_n \\ &= u_n \sum_{k=1}^{K} Re(d_{k,n} Z^c(\omega_{k,n})) + \varepsilon_n\end{aligned} \tag{1}$$

where

$$Z(\omega_{k,n}) = \begin{bmatrix} 0 & sin(\omega_{k,n}) & ... & sin((M-1)\omega_{k,n}) \\ 1 & cos(\omega_{k,n}) & ... & cos((M-1)\omega_{k,n}) \end{bmatrix}^T \tag{2}$$

$$Z^c(\omega_{k,n}) = [1, exp(i\omega_{k,n}), ..., exp(i(M-1)\omega_{k,n})]^T \tag{3}$$

$u_n \in \{0, 1\}$ determines whether n^{th} frame is voiced or not, $\omega_{k,n} = k\omega_{1,n}$ denotes the normalized radian frequency at k^{th} partial and n^{th} frame and K the total order of wave function, ε_n a Gaussian white noise vector respects $N(\mathbf{0}, \sigma_n^2 \mathbf{I})$, the real amplitude vector $a_{k,n} = [\alpha_{k,n}, \beta_{k,n}]^T$ and $d_{k,n} \in \mathbb{C}$ contain the amplitude and initial phase. The g-prior for order constraint and Viterbi decoder are proven effective to reduce the octave errors.

A dimension reduction approach allows a shorter window size [55] which is more appropriate to the pitch shift based techniques. Furthermore, comparing to pYIN, BNLS returns experimentally more consistent pitch contours under the voice activity smoothness, particularly in case of tremolo. Finally, since the pitch resolution is always limited by the searching bins, Fibonacci search for de-quantification [56] and the iterative methods [57] for each frame don't only take time but also lose the smoothness from Viterbi decoder. Alternatively, inspired by pYIN algorithm, we refine the pitch using parabolic interpolation on loss function of the BNLS.

2) Onset and Offset Detection

As another primary features in audio analysis usually fetch the critical structural prior, the onsets denotes the plucking points and offsets reflects the duration of a note event. Despite the voice activity detection (VAD) often bound with pitch estimators [59], the estimated boundaries still have a better a time-domain alignment and improve accuracy for subsequent tasks, e.g. note segmentation and playing technique detection. Energy peak detection generally has a sufficient performance for monophonic onset detection. Involved with simultaneously amplitude and pitch shift, the coarse onset detectors, e.g. SpecFlux [18], SuperFlux [60] and ComplexFlux [61], always follow with energy-based time alignment step [62].

Specially, the fake nail attack on the string brings a pair of adjacent peaks for a single tone, namely attack peak and natural transient, which respectively serve to note segmentation and velocity estimation. The above-mentioned onset detectors may introduce ambiguity of local peaks. To overcome this issues, we directly employ the log-energy envelope of high-pass filtered signal (See the bottom plot of Figure 6) and set parameters of boundary detectors as follows:

- **Peak-picking**: we use the peak picker in [60] and decrease the peak-picking parameters for the short intervals of peak pairs. The n^{th} frame with respect of Log-energy Envelope (LE) is selected as an potential onset if it fulfills 3 following conditions with the thresholds $\delta_1 = 25ms$, $\delta_2 = 25ms$, $\delta_3 = 25ms$, $\delta_4 = 25ms$, $\delta_5 = 0.025$ and $\delta_6 = 25ms$.

$$\begin{aligned} &1.\quad LE(n) = max(LE(n-\delta_1 : n+\delta_2)) \\ &2.\quad LE(n) \geq mean(LE(n-\delta_3 : n+\delta_4)) + \delta_5 \\ &3.\quad n - n_{PreviousNote} > \delta_6 \end{aligned} \tag{4}$$

- **Offset detection**: Given two types of correct onsets, a tone end until signal energy remaining 5% or next pluck starts.
- **Playing technique cases**: The offset of last pluck at the tail of tremolo interval respects above law, the detected onsets within tremolo interval needs to be eliminated. In case of pitch fluctuation techniques, the last one with 5% of local maximal energy or the end of vibrato is selected as offset. The pitch transition don't count into the offset of a note otherwise another pluck starts.

3) Note segmentation, Intensity, String and Fret Position

The note segments maintain the most direct connection to music scores. The pitch in unit of integer Musical Instrument Digital Interface (MIDI) value is converted from the median of pitch contours within note segment. The pitch discretization for pitch shift techniques like vibrato and sliding increase is necessary in our platform. To this end, the pitch2note algorithm in Tony which employs Hidden Markov Model (HMM) with 3 states, attack, stable and silence is chosen. In parallel, the intensity or the velocity in MIDI format denotes the loudness of a note embodied at natural transient. Notice that the intensity of the destination of sliding tones are set as 0. Given track index and note annotation, the string and fret position can be easily achieved through the presets.

B. PLAYING TECHNIQUE ANALYSIS AND ANNOTATION

Playing technique analysis generally concerns about both detection and parameter extraction within the detected ranges. In TEAS, the intervals and parameters of pitch fluctuation techniques (vibrato, trilling and bending), pitch transition techniques (sliding, slide in/out hammer-on, pull-off), tremolo and multi-track techniques (strumming, arpeggio) which occupy predominant portion of articulations, are automated to facilitate the annotation.

1) Pitch Fluctuation Analysis

Based on the pitch contour in continuous MIDI values, the pitch fluctuation parameter estimation is often viewed as a single harmonic inversion problem. In early work, Abeßer [106] proposes a note-level autocorrelation-based method which doesn't work to extent estimation and bending technique. The early short time non-parametric approaches like Time Fourier Transform (STFT) [63], [64] and cross-correlation of amplitude and frequency [65] have a related low performance on both detection and parameter estimation. Yang [66] utilizes a parametric damped sinusoidal estimator, Filter Diagonalization Method (FDM), that achieves a better fitting to pitch fluctuation compared to early works. Since the vibrato detection realized by hard thresholding or Decision Tree (DT) always relies on the instrumentation and playing styles, in TEAS we provide a single parameter method that could be steered by scroll between 0-1. Power Ratio (PR) [67], [68] widely used in VAD and its variant Power Difference (PD) to vibrato detection formulated with $\hat{\omega}_n$ as single-component version of the Eq. (9):

$$\hat{\omega}_n = max_{\omega_n} \, |\mathbf{y}_n^T Z^c(\omega_n)|^2 \tag{5}$$

$$PR_{per} = \frac{\mathbf{y}_n^T \hat{\mathbf{y}}_\mathbf{n}}{\mathbf{y}_n^T \mathbf{y}_n} = \frac{|\mathbf{y}_n^T Z^c(\hat{\omega}_n)|^2}{\mathbf{y}_n^T \mathbf{y}_n} > \eta_{per} \tag{6}$$

$$PD_{per} = |\mathbf{y}_n^T Z^c(\hat{\omega}_n)|^2 - \sigma_{vib}^2 > M\lambda_{per} \tag{7}$$

Under the hypothesis of the power subtraction method, the noise is independent to the signal and respects a Gaussian distribution with the standard deviation σ_{vib}.

$$\sigma_{vib}^2 = \mathbf{y}_n^T \mathbf{y}_n - Re(\hat{d}_n Z^c(\hat{\omega}_n))^T Re(\hat{d}_n Z^c(\hat{\omega}_n)) \tag{8}$$

Thus the PD_{per} can be computed with

$$PD_{per} = 2|\mathbf{y}_n^T Z^c(\hat{\omega}_n)|^2 - \mathbf{y}_n^T \mathbf{y}_n > M\lambda_{per} \tag{9}$$

the PR and PD on FDM can be constructed in same manner:

$$PR_{FDM} = \frac{\mathbf{y}_n^T Re(\hat{d}_n Z^c(\hat{\omega}_n^c))}{\mathbf{y}_n^T \mathbf{y}_n} > \eta_{FDM} \tag{10}$$

$$\begin{aligned} PD_{FDM} = 2Re(\hat{d}_n Z^c(\hat{\omega}_n^c))^T Re(\hat{d}_n Z^c(\hat{\omega}_n^c)) \\ - \mathbf{y}_n^T \mathbf{y}_n > M\lambda_{FDM} \end{aligned} \tag{11}$$

where the complex amplitude $\hat{d}_c$ and complex frequency $\hat{\omega}_n^c$ are estimated by the FDM. All the thresholds λ_{per}, λ_{FDM}, η_{per} and η_{FDM} mostly distribute between 0 and 1 that could be controlled by a scroll in practical application.

2) Pitch Transition Analysis

Wang proposes an SVM based model for 3 pitch evolution techniques, acciacatura (Short sliding ornament), portamento (Continuous pitch shift), glissando (Discrete pitch shift) for Chinese bamboo flute [28]. Yang also proposes a portamento detector in AVA [25] and parameter estimator [69] on the vibrato-free pitch contours for non-fretted instruments. Similarly to the 3-stated HMM for note segmentation, the down-steady-up states models the pitch evolution to detect the portamento. Logistic regression is used to estimate the parameters of note transition: inflection point, growth rate, antecedent and subsequent pitches. Strictly speaking as a fretted instrument, pipa has only glissando for note transition and slide in out around note boundary produced for pipa instrument and count in sliding technique. After smoothing on the vibrato-free pitch contour, we find the portamento modeling in AVA still works for pipa.

3) Tremolo Analysis

The tremolo in pipa, with 3 fingerings termed as wheel (Lun, 轮) on first or forth string, rolling (Gun, 滚) horizontally played by index finger, shaking (Yao, 摇) alternatively played by thumb and index mainly on middle two strings, is the most common technique in pipa articulations. This effect is physically constituted by a succession of rapid plucks on same string often accompanied with pitch shift techniques and dynamic variation. Hence, we parametrize the tremolo by the initial intensity, rate, number of plucks and fingering types within a segment of note (See Table 2). Different from the literal tremolo rates for violin [70], acoustic guitar [72] and flute [27] respectively range between 3-6 Hz, 3-8 Hz, 5-12 Hz, pipa has an minimal pluck interval down to 60 ms for synthesis [71] and 30 ms for real recording. Inspired by [72] which estimates the parameters after onset detection, we leverage the approach in III.B.2) the for plucking points and natural transients within each note. Given a series of corrected plucked points P_i within a tremolo note, the number of plucks $PN > 2$ defines the characteristic like the triple tremolo. The tremolo rate TR is computed by:

$$TR = \frac{\sum_{i=1}^{PN-1} P_{i+1} - P_i}{PN - 1} \tag{12}$$

4) Strumming Analysis

The automatic recognition of the strumming technique in TEAS platform are realized by the plucking points imported from each track. In our research, we propose a rule-based method which allows to aggregate the neighbouring plucks from orderly strings into a strum interval. The starting/ending strings and rate similarly defined to Eq. (12) within a interval are used to characterize the types of techniques. For example, the down tremble arpeggio represents the arpeggio performed from 3rd string to 1st one.

C. VISUALIZATION, CORRECTION, IMPORT/EXPORT AND VALIDATION

A multitrack tab (See plot (f) in Figure 7) incorporates an overall visualization for articulations, source separation (via KAMIR module), import/export modules of the whole project for quick global load. In order to better recover musicians' performance and simplify the validation from auditory perspective, we propose to export MIDI file as the load file of synthesizer like Ample China Pipa synthesizer [73]. The playing techniques are supported by key switch beyond that of pipa register, the more exquisite expressions like pitch fluctuation can be realized by the Continuous Controller (CC) channels. The MIDI module in Intelligent Sound Processing (ISP) toolbox [74] bridges the sophisticated annotations and MIDI files. Although the synthesized timbre and envelope sound different from that in our pipa but sufficiently natural for validation between synthesized audio and microphone recordings. From visual point of view, the tempo is indispensable for music score generation. Through the mean local autocorrelation approach in [75], the precise performance BPM (Beat per minute) can be computed, the initial BPM via the marked music score or musicians' guess. Different tempo of sections may occur so we split pieces by section.

Besides the fundamental event boundary fine-tuning, the correction for each module is described as follows:

- **Pitch**: Octave up/down is provided as the octave error take a large proportion of the errors. A pitch point or area could be selected to customize frequency or voice activity. The corrected notes help pitch curve clipping and padding in export module.
- **Onset/offset**: The selected points can be rectified to the closest peak to speed-up the correction.
- **Note segments**: A pair of adjacent onset peaks and estimated offset determine note attributes.
- **Pitch shifts**: Pitch fluctuation intervals are limited by corresponding note segments.
- **Other techniques**: The remaining techniques, like harmonic tone, twist tone or percussion non-identifiable through the string vibration, take up a small quantity. We manually annotate them as an additional playing technique sequence. The non-attack noise like that from friction is ignored.

IV. DATASET OVERVIEW

A. DATASET STATISTICS AND COMPARISON

Table 4 re-aggregates the related datasets aforementioned in Table 3 in terms of the polyphony, annotation types, playing techniques and basic materials. Among these datasets, GuitarSet [45] has pitch contours and discrete notes but no playing techniques. Multimodal Guitar [31] works on the fingerings in sense of that in I.B.2). Guitar playing technique dataset [20] focuses on monophonic recordings. The performed pieces in IDMT-SMT [21] sound stiff and lack

Datasets	Polyphony	Annotation types	Playing techniques	Basic materials
MAPS [33]	✓	Notes	×	Single notes, chords
MAESTRO v3 [112]	✓	Notes, velocity	Pedal	×
Yang [25]	×	Performance-level playing technique intervals and parameters	Vibrato/portamento	×
EEP [34]	✓	Bow MoCap data	Vibrato	×
CBF dataset [27], [28]	×	Performance-level playing technique intervals	Multiples	Playing techniques
Multimodal Guitar [31]	×	×	Fingerings	
Guitar playing technique [20]	×	×	Multiples	Single notes, playing techniques
IDMT-SMT [21]	✓	Notes/string, note-level playing technique	Multiples	Single notes, chords, playing techniques
URMP [32]	✓	Pitch/notes, note-level vibrato with parameters	Vibrato	×
Marovany zither [46]	✓	Pitch/notes	×	Single notes
GuitarSet [45]	✓	Pitch/notes/string, beats, chords	Strumming	×
CTIS dataset [76]	✓	×	Multiples	Single notes, playing techniques
PipaSet	✓	Pitch/notes/string/velocity, performance-level playing technique intervals and parameters	Multiples	Single notes, playing techniques

TABLE 4: Content characteristics of Datasets.

of annotated expression parameters. As the most massive dataset for Chinese traditional instruments, CTIS dataset [76] consists of more than 700 audio clips with playing techniques, but the absence of transcription annotation restricts its application. PipaSet has the relatively more comprehensive labels for transcription and expressive analysis tasks. During recordings, the musicians are asked to play music as naturally as possible so that the music score may not be completely followed. According to description in above sections, the detailed data types in PipaSet comprise:

- **Audio and string vibration signals:** Recordings from microphones and optical switches are saved in WAV format including the selected pieces in [77], [78] as well as the single-note scale from each string and playing technique examples as basic audio materials.
- **Video:** Videos from left, middle and right views pointed to the performers are encoded by MP4 with resolution of 1080p and sampling rate of 30 fps.
- **Sheet music:** The pipa numbered tonic-solfa scores are scanned and saved in PDF format.
- **Annotation:** All annotated features are manually corrected through audiovisual contrasts, then cross-validated to ensure the quality by musicians, finally saved as CSV files. MIDI files with an estimated tempo are converted from annotations, the pieces of multiple sections are split apart by the different tempo.

B. ACOUSTIC AND MUSIC FEATURES

In this section, we reveal common prior knowledge of particular instruments and non-eurogenetic music under the basic materials:

- **Fake nail noise:** Like the plucking peak for note segmentation in string vibration signal (See Figure 6), a more crisp attack noise triggered by the fake nails also occurs in audio recording.
- **Non-equal temperament:** From point of view of pipa construction, the fret positions often associated with pipa makers' craft probably produce an intrinsic pitch deviation from the equal temperament even though the open strings are well-tuned (See Figure 8).
- **Body response effect:** The modulation of body response affects the amplitude distribution of harmonic series. For example, the F0 amplitude of A3 tones may be seriously weaker than those of high order partials (See figure 9). Hence, the frequency of maximal amplitude and the interval between the zero-crossing points at the two ends of a waveform element may physically introduce the octave errors for pitch detection. Meanwhile, the sympathetic resonance between body and string may distort the pitch up to a semitone at tone tail of normal plucks around F5 tones ($\approx$ 700Hz). The approximate inharmonicity coefficients from 1st to 4th strings computed by the approach in [37]: 2.25e-5, 1.264e-4, 4.99e-5, 6.6e-4 which are thus negligible in most of applications irrelevant to strings.
- **Time-domain envelope:** Instead of T_{60} in [71], the values of T_{26} corresponding to the duration from natural transients to offsets of 5% of maximal energy remaining from 1st to 4th open strings are 0.546, 1.318, 1.4834, 1.0674 which don't always decrease with the pitch.
- **Tempo and rubato:** Except different tempos in music sections like those in **Ambush from Ten Sides**, self-interpretation [79] often alters the performance speed in Chinese music.

C. APPLICATIONS OF THE DATASET

In this section, we enumerate a variety of conventional and promising emerging tasks empowered by our versatile dataset from single-modal and multi-modal perspectives. First, the single-modal and intra-modal applications are present as follows:

- **Annotation:** The expressive music and performance generation [80]–[82] appears barely studied with complex articulations of plucked string instruments and

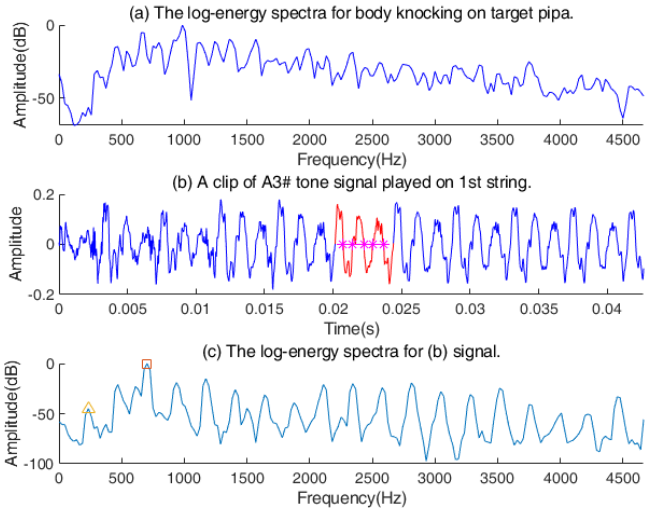


FIGURE 8: (a) Focus on 500-1000 Hz. (b) Red curve denotes a single waveform element in time domain and 5 zeros crossing points within. (c) The triangle and correspond to the F0 and F3

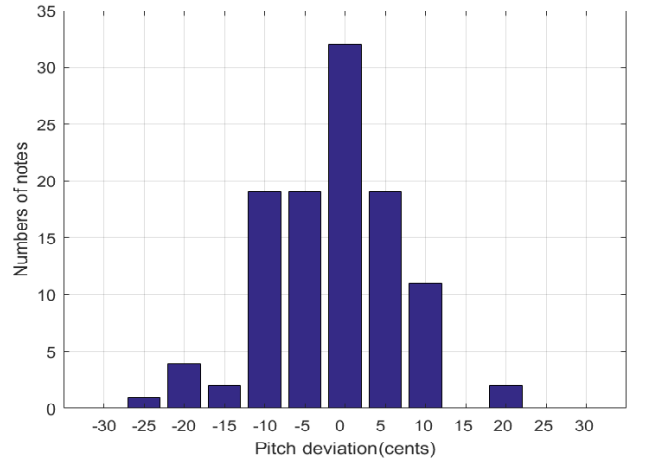


FIGURE 9: The approximate pitch deviation distribution of 109 recorded notes in terms of equal temperament for pipa used in our work.

promising for the future. The higher music representations like chord and key can be recognized by the musicological toolkits like Music21 [83]. Since the repertoires originate from famous and classic collections so that the external information like the key, style, emotion, even aesthetic comments is easy to fetch. The annotations have potentials to the tasks beyond the transcription and hierarchical multi-tasks [84]–[87].

- **Audio:** As the fundamental MIR tasks for polyphonic expressive music, multi-pitch estimation [88], [89] and playing technique analysis [21], [90] are the most direct applications to the dataset. Higher-level music feature analysis can be realized under extended annotation. The expressive annotation-synthesis frameworks can be explored with the basic material [91]–[93].
- **Video:** Multi-view shots allows the visual 3D construction for robust hand pose estimation [94] and performance analysis [95]. The pitch, fret, string position on fingerboard [96], [97], vibrato [98] could be more directly perceived. The colored fake nails help to mark the finger motion and provides a way for fingering analysis.
- **Music sheet:** The works on Optical Character Recognition (OCR) [99], [100] for special notation or called Optical Music Recognition (OMR) and the score-performance alignment for non-eurogenetic music [101] are new topics. The analysis and generation of expressions including tempo and dynamics could be realized between score and annotation including the parameters of articulations [102].

Multimodality analysis has become increasingly popular in various fields [103], [104]. Video [105] and score [106], [107] based works are also proposed to deal with polyphonic scenarios. The audio-visual generation for fingering and gestures [108], sight-to-sound [109] for the motion-based music generation, audio-sheet correspondence analysis [110] and cross-modal query [111] can be realized through PipaSet as multimodal transformations.

V. CONCLUSION AND FUTURE WORK

To the best of our knowledge, PipaSet is the first annotated dataset for transcription and expressive analysis dedicated to Chinese traditional instrument pipa. The main contribution of this paper is to present a complete creation procedure of prototype dataset including multimodal recording and multitask annotation. The motivation of this work is to promote the development of the research in ethnomusicology and polyphonic string instruments with expressive features. Licensed under GNU general public license, the preview of datasets and the beta version of TEAS platform are open-sourced respectively on Zenodo[1] on Github[2].

Although a series of applications are introduced in section IV, some limitations in our dataset could be improved from different aspects in the future. The annotation for gestures and fingerings, e.g. termination position, plucking finger and position, inclination, can be captured from video or the other sensors like foil gauge mounted on fake nail for the plucking deflection, Time of Flight (ToF) camera for depth estimation, etc. Finally, we will extend this dataset using TEAS platform to cover more pieces of music, pipa models, players even the other plucked string instruments in the future.

ACKNOWLEDGEMENT

We greatly thank to Yuanbo Tang, M.S. from SEU, for his constructive contribution to the device deployment. We are also grateful to the virtuosic performance of Xiaoyue Zhang from Nanjing XiaoZhuang University.

[1]Website's coming soon
[2]Website's coming soon

REFERENCES

[1] J James Anderson Moorer. On the segmentation and analysis of continuous musical sound by digital computer. Ph.D. dissertation, CCRMA, Standford University, 1975.

[2] Music Information Retrieval Evaluation eXchange (MIREX). https://www.music-mir.org/mirex/wiki/MIREX_HOME. Retrieved date last viewed Dec 12, 2019.

[3] Anssi P. Klapuri. Automatic music transcription as we know it today. Journal of New Music Research, 33(3):269–282, 2004.

[4] E. Benetos, S. Dixon, Z. Duan and S. Ewert, "Automatic Music Transcription: An Overview," in IEEE Signal Processing Magazine, vol. 36, no. 1, pp. 20-30, Jan. 2019

[5] Adrien Ycart, Lele Liu, Emmanouil Benetos and Marcus Pearce. Investigating the Perceptual Validity of Evaluation Metrics for Automatic Piano Music Transcription. Transactions of the International Society for Music Information Retrieval, 3(1), pp. 68–81.

[6] Valentin Emiya, Roland Badeau, and Bertrand David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Transactions on Audio Speech & Language Processing, 18(6):1643–1654, 2010.

[7] A. M. Barbancho, A. Klapuri, L. J. Tardon, and I. Barbancho. Automatic transcription of guitar chords and fingering from audio. IEEE Transactions on Audio Speech & Language Processing, 20(3):915–921, 2012.

[8] Miryala, Sai Sumanth, et al. "Automatically Identifying Vocal Expressions for Music Transcription." International Society of Music Information Retrieval (ISMIR), 2013.

[9] E. Benetos and A. Holzapfel. Automatic transcription of turkish microtonal music. Journal of the Acoustical Society of America, 138:2118–2130, Oct 2015.

[10] Dorian Cazau, YuanchengWang, Marc Chemillier, and Olivier Adam. An automatic music transcription system dedicated to the repertoires of the marovany zither. Journal of New Music Research, 45(4):1–18, 2016.

[11] George Tzanetakis. Computational ethnomusicology: A music information retrieval perspective. International Computer Music Conference (ICMC), 2014.

[12] Andre Holzapfel, Emmanouil Benetos. Automatic Music Transcription and Ethnomusicology: a User Study. In ISMIR, pages 678-684, 2019.

[13] Ruijin Huang, Sturm Bob and Holzapfel Andre .De-centering the west: East asian philosophies and the ethics of applying artificial intelligence to music. International Society of Music Information Retrieval (ISMIR), 2021.

[14] Dorottya Fabian, Renee Timmers, Emery Schubert. Expressiveness in Music Performance: Empirical Approaches across Styles and Cultures. Oxford Scholarship Online.

[15] Meinard Müller, Andreas Arzt, Stefan Balke, Matthias Dorfer, Gerhard Widmer. Cross-Modal Music Retrieval and Applications: An Overview of Key Methodologies. IEEE Signal Processing Magazine (Volume: 36, Issue: 1, Jan. 2019).

[16] Marc Leman, Luc Nijs, and Nicola Di Stefano, "On the role of the hand in the expression of music," in The Hand, pp. 175–192. Springer, 2017.

[17] Mikel Gainza and Eugene Coyle, "Automating ornamentation transcription," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2007, vol. 1, pp. I–69.

[18] Loïc Reboursière, Otso Lähdeoja, Thomas Drugman, Stéphane Dupont, Cécile Picardlimpens Limpens, Nicola Riche. Left and right-hand guitar playing techniques detection. International Conference on New Interfaces for Musical Expression, 2013.

[19] Yuan-Ping Chen, Li Su, Yi-Hsuan Yang, Electric guitar playing technique detection in real world recordings based on F0 sequence pattern recognition. 16th International Society for Music Information Retrieval Conference, 2015.

[20] Li Su, Li-Fan Yu, and Yi-Hsuan Yang. Sparse cepstral, phase codes for guitar playing technique classification. In ISMIR, pages 9–14, 2014.

[21] Christian Kehling, Jakob Abeßer, Christian Dittmar, and Gerald Schuller. Automatic tablature transcription of electric guitar recordings by estimation of score- and instrument-related parameters. In DAFx, pages 219–226, 2014.

[22] Ting-Wei Su, Yuan-Ping Chen, Li Su, and yi-hsuan Yang, "Tent: Technique-embedded note tracking for real-world guitar solo recordings," Transactions of the International Society for Music Information Retrieval, vol. 2, pp. 15–28, 07 2019.

[23] Jakob Abeßer and Gerald Schuller, Instrument-Centered Music Transcription of Solo Bass Guitar Recordings, IEEE/ACM Transaction on Audio, Speech, And Language Processing, Vol. 25, No. 9, September 2017.

[24] Barbancho I, de la Bandera C, Barbancho A M, et al. Transcription and expressiveness detection system for violin music. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. 2009: 189–192.

[25] Yang, Luwei, Sayid-Khalid Rajab, and Elaine Chew. "AVA: A Graphical User Interface for Automatic Vibrato and Portamento Detection and Analysis." International Society for Music Information Retrieval Conference 2016.

[26] A. B. Kruger & J. P. Jacobs. Playing technique classification for bowed string instruments from raw audio, Journal of New Music Research, 49:4, 320-333, 2020.

[27] Changhong Wang, Emmanouil Benetos, Vincent Lostanlen, Elaine Chew. Adaptive time-frequency scattering for modulation recognition in music signal. International Society for Music Information Retrieval Conference (ISMIR), Delft, The Netherlands, 2019.

[28] Changhong Wang, Vincent Lostanlen, Emmanouil Benetos, Elaine Chew. Playing technique recognition by joint time-frequency scattering. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019. Dataset Available: http://c4dm.eecs.qmul.ac.uk/CBFdataset.html. Retrieved date last viewed Dec 12, 2020.

[29] F. Simonetta, S. Ntalampiras and F. Avanzini, "Multimodal Music Information Processing and Retrieval: Survey and Future Challenges," 2019 International Workshop on Multilayer Music Representation and Processing (MMRP), 2019, pp. 10-18.

[30] L. Zhonghua, "Multimodal music information retrieval: From content analysis to multimodal fusion," Ph.D. dissertation, 2013.

[31] Alfonso Perez-Carrillo, Josep-Lluis Arcos, and Marcelo Wanderley. Estimation of Guitar Fingering and Plucking Controls based on Multimodal Analysis of Motion, Audio and Musical Score. Proc. of the 11th International Symposium on CMMR, Plymouth, UK, June 16-19, 2015.

[32] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan and Gaurav Sharma. Creating a Multitrack Classical Music Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications. IEEE Transactions on multimedia, vol. 21, no. 2, February 2019.

[33] Valentin Emiya, Roland Badeau, and Bertrand David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. IEEE Transactions on Audio, Speech, and Language Processing, 18(6):1643–1654, 2010.

[34] M. Marchini, R. Ramirez, P. Papiotis, and E. Maestre, "The sense of ensemble: A machine learning approach to expressive performance modelling in string quartets," Journal of New Music Ressearch, vol. 43, no. 3, pp. 303–317,2014.

[35] No. 561 product of the Shanghai No.1 National Musical instrument factory. Pipa used in this paper. http://shop.dunhuangguoyue.com/product-503.html. Retrieved date last viewed Dec 12, 2019.

[36] T. D. Rossing, The science of string instruments, 1st ed. Springer, 2010.

[37] Jacob Møller Hjerrild and Mads Græsbøll Christensen, Estimation of guitar string, fret and plucking position using parametric pitch estimation.IEEE, International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019), May 12-17, Brighton, United Kingdom.

[38] Christian Dittmar, Andreas Mannchen, and Jakob Abeber, "Real-time guitar string detection for music education software," in 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), 2013.

[39] Yongping Zhuang. Pipa Handbook. Shanghai Music Press, 2001.

[40] ZPshen pipa fonts.http://www.ppfeng.com/forum.php?mod=viewthread&tid=53853. Retrieved date last viewed 10, Oct, 2021.

[41] S. Feng. Some acoustical measurements on Chinese musical instrument pipa. Journal Acous. Soc. Amer., Vol 75, no. 2, pp. 599-602, 1984.

[42] Eric F Clarke. Listening to Performance. In John Rink, editor, Musical Performance — A Guide to Understanding. Cambridge University Press, Cambridge, 2002.

[43] Li Su and Yi-Hsuan Yang, "Escaping from the Abyss of Manual Annotation: New Methodology of Building Polyphonic Datasets for Automatic Music Transcription," in Int. Symp. Computer Music Multidisciplinary Research (CMMR), June 2015.

[44] Iñigo Angulo, Sergio Giraldo and Rafael Ramirez. Hexaphonic guitar transcription and visualization. In Proceedings of the Second International Conference on Technologies for Music Notation and Representation (TENOR), 2016.

[45] Qingyang Xi, Rachel Bittner, Johan Pauwels, Xuzhou Ye, Juan Bello. GuitarSet: A Dataset for Guitar Transcription. 19th International Society for Music Information Retrieval Conference, Paris, France, 2018.

[46] Dorian Cazau, Olivier Adam and Marc Chemillier. Information retrieval of Marovany zither music with an original optical-based system. Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland, September 2-5, 2013.

[47] sE Electronics sE8 Small Diaphragm Microphone. https://www.seelectronics.com/se8-mic. Retrieved date last viewed 10, March, 2021.

[48] RME fireface UFX sound card. http://babyface.rme-audio.de/. Retrieved date last viewed 10, March, 2021.

[49] Thomas Pratzlich, Rachel M. Bittner, Antoine Liutkus, and Meinard Muller. Kernel additive modeling for interference reduction in multi-channel music recording. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, May 2015. Code available: https://members.loria.fr/ALiutkus/kamir/. Retrieved date last viewed 10, March, 2021.

[50] Y. Ephraim and D. Malah, "Speech enhancement using a minmum mean-square error log-spectral amplitude estimator," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 33, pp. 443–445, 1985.

[51] Matthias Mauch, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jiajie Dai, Juan Bello, Simon Dixon. "Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency." International Conference on Technologies for Music Notation Representation 2015.

[52] Celemony Melodyne Studio 5. Available: https://www.celemony.com/en/melodyne/new-in-melodyne-5.

[53] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. IEEE , International Conference on Acoustics, Speech and Signal Processing(ICASSP 2014), pages 659–663. IEEE, 2014.

[54] Shi Liming et al. "Robust Bayesian Pitch Tracking Based on the Harmonic Model." IEEE/ACM Transactions on Audio, Speech, and Language Processing 27.11(2019):1737-1751.

[55] Nielsen, J. K. , et al. "Grid size selection for nonlinear least-squares optimization in spectral estimation and array processing." 2016 24th European Signal Processing Conference (EUSIPCO) IEEE, 2016.

[56] A. Antoniou and W.-S. Lu, Practical Optimization: Algorithms and Engineering Applications. Springer, Mar. 2007.

[57] Morfi, V. , G. De Gottex , and A. Mouchtaris. "A computationally efficient refinement of the fundamental frequency estimate for the Adaptive Harmonic Model." IEEE (2014).

[58] Sibelius Music notation software. https://www.avid.com/sibelius. Retrieved date last viewed 10, Jan, 2021.

[59] Thomas Drugman, Yannis Stylianou,Yusuke Kida and Masami Akamine. "Voice Activity Detection: Merging Source and Filter-based Information." IEEE Signal Processing Letters 23.2(2016):252-256.

[60] S. B¨ock and G. Widmer. Maximum filter vibrato suppression for onset detection. In Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland, September 2013.

[61] Böck, Sebastian, and G. Widmer . "Local Group Delay Based Vibrato and Tremolo Suppression for Onset Detection. " International Society for Music Information Retrieval Conference 2013.

[62] Jehan, Tristan. "Creating music by listening" Doctoral dissertation Massachusetts Institute of Technology, 2005.

[63] P. Herrera and J. Bonada, "Vibrato extraction and parameterization in the spectral modeling synthesis framework," in Proceedings of the Digital Audio Effects Workshop (DAFX98), 1998, vol. 99.

[64] J. Ventura, R. Sousa, and A. Ferreira, "Accurate analysis and visual feedback of vibrato in singing," in 2012 5th International Symposium on Communications, Control and Signal Processing, 2012, pp. 1–6.

[65] H. Von Coler and A. Roebel, "Vibrato detection using cross correlation between temporal energy and fundamental frequency," in Audio Engineering Society Convention 131, 2011.

[66] L. Yang, K. Z. Rajab, and E. Chew, "The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation," Journal of Mathematics and Music, vol. 11, no. 1, pp. 42–60, 2017.

[67] Jongseo Sohn and Wonyong Sung, A voice activity detector employing soft decision based noise spectrum adaptation. in Proceeding IEEE International Conference Acoustics Speech and signal processing(ICASSP), 1998

[68] Etan Fisher, Joseph Tabrikian and Shlomo Dubnov. Generalized Likelihood Ratio Test for Voiced-Unvoiced Decision in Noisy Speech Using the Harmonic Model.IEEE Transactions on audio, speech, and language processing, VOL. 14, NO. 2, March 2006.

[69] L. Yang, E. Chew, and K. Z. Rajab, "Logistic Modeling of Note Transitions," in Mathematics and Computation in Music. Springer, 2015, pp. 161–172.

[70] Isabel Barbancho, Cristina de la Bandera, Ana M Barbancho, and Lorenzo J Tardon, "Transcription and expressiveness detection system for violin music," in 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2009, pp. 189–192.

[71] Yi-Huei Chen and Chih-Fang Huang, "Sound synthesis of the pipa based on computed timbre analysis and physical modeling," IEEE Journal of Selected Topic in Signal Processing. vol. 5, no. 6, pp. 1170–1179, 2011.

[72] Sergio Freire and Lucas Nezio, "Study of the tremolo technique on the acoustic guitar: Experimental setup and preliminary results on regularity," in Proc. Int. Conf. Sound and Music Computing, Stockholm, 2013, pp. 329–334.

[73] Ample China Pipa. http://www.amplesound.net/en/pro-pd.asp?id=30. Retrieved date last viewed Dec 12, 2019.

[74] Intelligent Sound Processing (ISP) toolbox. https://github.com/murtzdhulz/AudioToMidiConverter/tree/master/Matlab-Code/isptoolbox_20100209. Retrieved date last viewed 3, May, 2021.

[75] Grosche, Peter, Meinard Müller, and Frank Kurth. "Cyclic tempogram - A mid-level tempo representation for music signals." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010.

[76] Liang X, Li Z, Liu J, et al. Constructing a multimedia Chinese musical instrument database. Proceedings of the 6th Conference on Sound and Music Technology (CSMT). Springer, Singapore, 2019: 53-60. https://zenodo.org/record/5676893.YZTz8ciEzcs. Revisited in Nov. 2021.

[77] Grade examination committee of China Conservatory of musicNational general textbook for social art grade examination for Pipa, China Youth Publishing House, Jan, 2013.

[78] Pipa Professional Committee of Shanghai Musicians Association. The collection for Pipa grade examination of China. Shanghai Music Publishing House (SMPH),Dec 2012.

[79] Meinard Muller, Verena Konz, Andi Scharfstein, Sebastian Ewert, Michael Clausen. Towards automated extraction of tempo parameters from expressive recordings. 10th International Society for Music Information Retrieval Conference (ISMIR), 2009.

[80] Alexander Lerch, Claire Arthur, Ashis Pati, Siddharth Gururani. Music Performance Analysis: A Survey. 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 2019.

[81] José Maria Simões, Penousal Machado. "Deep Learning for Expressive Music Generation." ARTECH 2019: 9th International Conference on Digital and Interactive Arts 2019.

[82] Carnovalini, F. , and A. Roda . "A Multilayered Approach to Automatic Music Generation and Expressive Performance." 2019 International Workshop on Multilayer Music Representation and Processing (MMRP) 2019.

[83] Cuthbert, M., Ariza, C.: music21: a toolkit for computer-aided musicology and symbolic music data. In: Proceedings of the International Symposium on Music Information Retrieval, p. 63742 (2010). http://web.mit.edu/music21/doc/, Retrieved date last viewed 1 Jan 2022.

[84] Fred Lerdahl and Ray Jackendoff. An Overview of Hierarchical Structure in Music. Music Perception: An Interdisciplinary Journal, 1(2):229–252, 1983.

[85] Choi Keunwoo, György Fazekas, Kyunghyun Cho and Mark Sandler. "A Tutorial on Deep Learning for Music Information Retrieval." arXiv:1709.04396v1 [cs.CV] 13 Sep 2017.

[86] B. McFee, O. Nieto, and J. P. Bello. Hierarchical evaluation of segment boundary detection. International Society for Music Information Retrieval 2015.

[87] Renato Panda, Ricardo Malheiro and Rui Pedro Paiva. Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing Vol. 9, 2018.

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

[89] Rachel M Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan P Bello. Deep salience representations for f0 estimation in polyphonic music. In Proceedings of the 18th International Conference on Music Information Retrieval (ISMIR), Suzhou, China, 2017.

[90] Jonathan Driedger, Stefan Balke, Sebastian Ewert, Meinard Muller. Template-Based Vibrato Analysis in Complex Music Signals. in Proceedings of the International Society for Music Information Retrieval Conference (ISMIR) 2016: 239-245.

[91] Raymond Vincent Migneco. Analysis and synthesis of expressive guitar performance. Drexel University, Doctor of Philosophy Thesis.

[92] Justin Salamon, Rachel M Bittner, Jordi Bonada, Juan Jose Bosch Vicente, Emilia Gomez Gutierrez, and Juan P Bello. An analysis/synthesis framework for automatic f0 annotation of multitrack datasets. In 18th International Society of Music Information Retrieval (ISMIR) Conference, October 2017.

[93] Yusong Wu, Ethan Manilow, Yi Deng, Rigel Swavely, Kyle Kastner, Tim Cooijmans, Aaron Courville, Cheng-Zhi Anna Huang, Jesse Engel. MIDI-DDSP: Detailed Control of Musical Performance via Hierarchical Modeling. The Tenth International Conference on Learning Representations, ICLR 2022.

[94] Leyla Khaleghi,Alireza Sepas Moghaddam,Joshua Marshall,Ali Etemad. Multi-View Video-Based 3D Hand Pose Estimation. arXiv:2109.11747v1 [cs.CV] 24 Sep 2021.

[95] Y Kwon, et al. "Neural Human Performer: Learning Generalizable Radiance Fields for Human Performance Rendering." Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS), 2021.

[96] A. Sophia Koepke, Olivia Wiles, and Andrew Zisserman, "Visual pitch estimation," in Sound and Music Computing Conference, 2019.

[97] M. Paleari, B. Huet, A. Schutz and D. Slock, "A multimodal approach to music transcription," 2008 15th IEEE International Conference on Image Processing, 2008, pp. 93-96.

[98] Bochen Li, Karthik Dinesh, Gaurav Sharma, and Zhiyao Duan, "Video-based Vibrato Detection and Analysis for Polyphonic String Music", in Proc. International Society for Music Information Retrieval (ISMIR), 2017.

[99] Ana Rebelo, Ichiro Fujinaga, Filipe Paszkiewicz, Andre R. S. Marcal, Carlos Guedes, Jaime S. Cardoso. Optical music recognition: state-of-the-art and open issues. International Journal of Multimedia Information Retrieval 2012 1.3:173-190.

[100] Jorge Calvo-Zaragoza, Jan Hajic Jr. and Alexander Pacha, Understanding Optical Music Recognition, arXiv:1908.03608v2 [cs.CV] 14 Aug 2019

[101] Josquintab: A dataset for content-based computational analysis of music in lute tablature. In 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 2019.

[102] Carlos Cancino-Chacón, Katerina Kosta and Maarten Grachten. Computational Modeling of Musical Expression Perspectives, Datasets, Analysis and Generation. 20th International Society for Music Information Retrieval Conference (ISMIR), 2019.

[103] Wang, K., Yin, Q., Wang, W., Wu, S., and Wang, L., "A Comprehensive Survey on Cross-modal Retrieval", arXiv:1607.06215, arXiv e-prints, 2016.

[104] Baltrusaitis, Tadas , C. Ahuja , and L. P. Morency . "Multimodal Machine Learning: A Survey and Taxonomy." IEEE Transactions on Pattern Analysis Machine Intelligence PP.99(2017):1-1.

[105] Jose Ventura, Ricardo Sousa, and Anibal Ferreira. Accurate analysis and visual feedback of vibrato in singing. In Proc. Intl. Symposium on Communications Control and Signal Process. (ISCCSP), pages 1–6.IEEE, 2012.

[106] Jakob Abeßer, Estefania Cano, Klaus Frieler, Martin Pfleiderer, and Wolf-Georg Zaddach. Score-informed analysis of intonation and pitch modulation in jazz solos. In Proc. Intl. Society for Music Information Retrieval (ISMIR), pages 823–829, 2015.

[107] P.-C. Li, L. Su, Y.-H. Yang, and A. W. Y. Su, "Analysis of expressive musical terms in violin using score-informed and expression-based audio features," in Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, M´alaga, Spain, October 26-30, 2015, M. M¨uller and F. Wiering, Eds., 2015, pp. 809–815.

[108] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio visual generation," in Proc. ACM Thematic Workshops of Multimedia, 2017, pp. 349–357.

[109] Koepke, A. Sophia , et al. "Sight to Sound: An End-to-End Approach for Visual Piano Transcription." ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE, 2020.

[110] M. Dorfer, A. Arzt, and G. Widmer, "Learning audio-sheet music correspondences for score identification and offline alignment," in Proceedings of the International Society for Music Information Retrieval Conference (ISMIR), Suzhou, China, 2017, pp. 115–122.

[111] Bochen Li, Aparna Kumar. "Query by Video: Cross-modal Music Retrieval", 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 2019.

[112] Hawthorne, C. , et al. "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset." International Conference on Learning Representations, ICLR 2019. https://magenta.tensorflow.org/datasets/maestro. Retrieved date last viewed 30, Mars, 2021.

YUANCHENG WANG was born in Wuhan, China in 1989. He received double B.S. degrees in applied mathematics and French language from University of Wuhan and University of La Rochelle in 2011, double M.S. degrees in statistics from Sorbonne University and University of Reims in 2014. He is currently pursuing Ph.D. student in Music Signal Analysis supervised by Prof. Qiao Wang at Southeast University, Nanjing, China.
In 2015, he participated in Prof. Olivier Adam's group as research assistant in Lutherie Acoustique Musique (LAM) Laboratory, CNRS, France. He published 8 articles in signal analysis, machine learning and deep learning approaches. His recent research interest focus on audio-visual analysis of Chinese folk music. National and provincial prizes in art and Alibaba Tianchi AI competition prize on Computer Vision are awarded.

YUYANG JING was born in NanjingChina in 1991. He is a lecturer at Department of recording art, Nanjing University of the Arts (NUA). He received B.A. degree, M.A. degree in Composition and Recording Engineering and Ph.D. in Digital Media Art from NUA respectively in the years of 2014, 2017 and 2020. His research fields are computer composition, MIDI technology and expression of music, recording engineering, sampling technology, artificial intelligence.

WEI WEI is a lecturer at Nanjing Xiaozhuang University. She received her B.A. degree in 2002 and M.A. degree in 2006 in musicology from Nanjing University of the Arts. Her research focused on traditional Chinese musical instrument (Pipa) performance and education, including the application of diversified modern skills in playing traditional Chinese instrument. She has published 20 professional articles on core national and provincial-level periodicals. The pipa performing group under her coach has won prizes in various competitions.

DORIAN CAZAU received the Ph.D. degree in acoustics and signal processing from Pierre-and Marie-Curie University, Paris, France, in 2015. In 2015, he held a post-doctoral position at ENSTA Bretagne, Brest, France, then a French post-Doctoral Researcher with the Institut Mines Telecom Atlantique, Lab-STICC, UMR CNRS 6285, where he is involved in acoustic data analysis. He is involved in computer analysis of audio signals, combining methods from physical acoustics, signal processing, statistical modeling, and machine learning. His research interests include different interdisciplinary application fields, including environmental acoustics, bioacoustics, and acoustical oceanography.

OLIVIER ADAM was born in Paris, France, in 1967. He received M.S. degree in applied physics and did his PhD thesis on artificial neural networks. He is currently Professor at Sorbonne University, Paris, France. Specialist in signal processing and acoustics, his main research topic is the analysis of sounds emitted by cetacean species in order to describe their behaviors, interactions and habitats. He was also involved in automatic Music transcription, and was interested by analysing traditional music from Malagasy zithers. He published more than 50 research articles in scientific revues

QIAO WANG received the BS, MS, and PhD degrees in mathematics from Wuhan University, China, in 1988, 1994, and 1997, respectively. He is currently professor of the School of Information Science and Engineering, and ShingTung Yau Center, Southeast University, China. He joined the School of Information Science and Engineering, Southeast University, Nanjing, China, in 1997, and was appointed as associate professor, in 1999, then full professor, in 2001. He was a visiting scientist with Harvard University, Cambridge, Massachusetts from January 2003 to January 2004. His main research interests include applied harmonic analysis and information theory, data analysis and image analysis in the areas of urban science, urban planning and design, clinical medicine and public health, and sports analysis. He was awarded Excellence Design by the International Society of the Built Environment, in 2019, and Best Editor Award of ICT Express, 2019.