

Article

Not peer-reviewed version

---

# Towards High-Resolution Multipitch Estimation for Expressive Pipa Music Annotation Using Morphological Analysis of Photoelectric Signals

---

[Yuancheng Wang](#), [Xuanzhe Li](#), [Yunxiao Zhang](#), [Qiao Wang](#)\*

Posted Date: 10 December 2024

doi: 10.20944/preprints202412.0826.v1

Keywords: Automatic music transcription (AMT); pipa; playing techniques; AM-FM signal; photoelectric signal; morphological analysis




Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

## Article

# Towards High-Resolution Multipitch Estimation for Expressive Pipa Music Annotation Using Morphological Analysis of Photoelectric Signals

Yuancheng Wang <sup>1</sup> , Xuanzhe Li <sup>2</sup>, Yunxiao Zhang<sup>2</sup> and Qiao Wang <sup>1,\*</sup>

<sup>1</sup> School of Information Science and Engineering, Southeast University, 211102, Nanjing, China

<sup>2</sup> School of Artificial Intelligence, Southeast University, 211102, Nanjing, China

\* Correspondence: qiaowang@seu.edu.cn

**Abstract:** The prosodic music signal physically exhibits non-linearity and non-stationarity driven by variations in pitch and amplitude, which pose a great challenge to pitch estimation under complex playing techniques. To address this issue, we revisit different single-pitch annotation algorithms and processes on the string displacement signal captured by optical sensors mounted on pipa, a Chinese traditional plucked string instrument with a high diversity of playing techniques. This signal demonstrates an arched form momentarily deviating from origin at plucks, which facilitates accurate boundary detection and avoids the impact of spurious energy peaks arose by pitch-shift playing techniques within vibration areas. Then we develop a novel amplitude-invariant sparse time-frequency representation termed by continuous time-period mapping (CTPM), which allows to extract pitch curves without signal shaping, even in cases where multiple oscillations occur within a single period. Playing techniques mixed of pitch-shift and tremolo are also covered by our proposed process. Evaluated on 4 renowned expressive pipa music pieces of varying difficulty levels, our proposed fully time-domain onset detectors outperform the 4 commonly-used short-time methods particularly during rapid-plucking and tremolo. Zero-crossing-based pitch estimator achieved an overwhelming computational efficiency and a performance comparable to short-time methods if provided the correct boundaries.

**Keywords:** Automatic music transcription (AMT); pipa; playing techniques; AM-FM signal; photoelectric signal; morphological analysis

## 1. Introduction

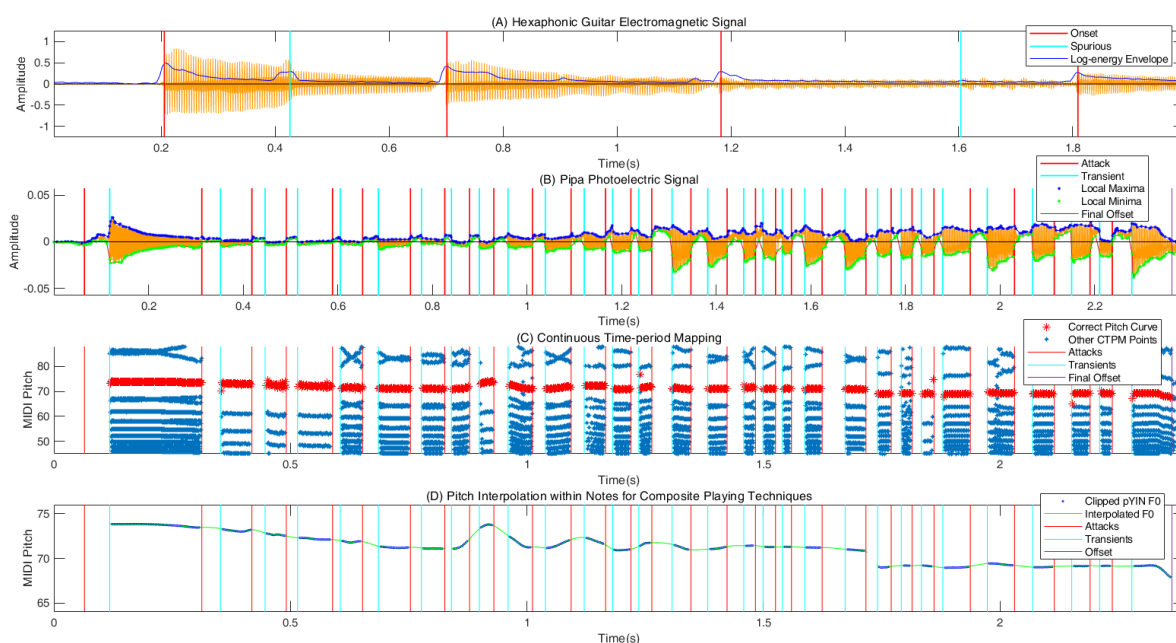
Performers manipulate the sound properties, including frequency (pitch), timing, amplitude, and timbre (harmonic spectrum) above and beyond the pitch and duration determined by composers to enrich the musical expressiveness and human perception [1]. The playing techniques like tremolo and vibrato often accompany with intensive strength and pitch variations corresponding to the amplitude modulation (AM) and frequency modulation (FM) of the vibration signal from signal processing perspective, which are widely present in world music. In practice, some monophonic pitch annotation tools like Tony [2], employing the short-time approaches under the local stationarity hypothesis, doesn't perform well on complex playing techniques, over half of the notes require manual correction [3,4], not to mention in polyphonic settings [5,6].

As a traditional Chinese plucked string instrument, pipa boasts a long history and profound cultural heritage, with its origins tracing back to around the 2nd century BC. The earliest form of the may have originated from instruments in Central Asia and the Western Regions, such as Persia or India. This instrument gradually made their way into China through trade and cultural exchanges along the Silk Road, evolving and localizing over time to become the pipa we are familiar with today. The modern pipa has 4-string, 6 ledges and 24 frets which is performed with celluloid-made fakenails wore on each right finger. More than 60 playing techniques has been developed by the musicians[7] and the mixture of these techniques greatly enhances the expressiveness of music.

Automatic pipa transcription (APT), as a subtask of AMT dedicated to pipa, aims to build a multi-pitch estimation system capable of handling complex playing techniques, that extract features

such as string number, F0 (fundamental frequency), boundaries (attack/transient/offset) and notes. In order to get the high-resolution multi-pitch annotation on plucked string instruments, Xi [8] and Wang [9] have mounted respectively electromagnetic pickups on guitar and optical switches on pipa, the multipitch estimation task is thus converted into the single pitch estimation on different strings. Although these two works provide a process with three steps, i.e. source separation to reduce the mutual resonance among the strings, boundary detection and pitch estimation, all steps are based on short-time algorithms and neither of these works investigate the annotation efficiency on boundary and pitch.

From time-domain perspective, the electromagnetic signal has a waveform similar to that of audio and the photoelectric one directly captures the string displacement which has different morphological features. Figure 1(A) shows a clip of the electromagnetic signal with two sliding tones that bring two spurious energy peaks located around 0.4 and 1.6 seconds (cyan lines). Concretely, the plucks severely bend the strings and make the photoelectric signal behave like arched form as shown in Figure 1(B). This feature doesn't exist in audio and electromagnetic signal and provides a new indicator to extract the attack and transient points that circumvents the amplitude impacts from pitch shift techniques like vibrato and sliding. Given the accuracy boundaries, we could focus on the pitch estimation of the non-stationary vibration signal within each clip. Besides, the pitch period extraction directly from complex-structured waveform without any signal shaping is still a challenging work.



**Figure 1.** (A) and (B) Signals from electromagnetic guitar pickup and optical pipa pickup. (B) The onset detection using extrema-based envelopes. (C) Continuous time period mapping (CTPM) and pitch curves for photoelectric signal from plot (B) that contains mixed playing techniques. (D) Postprocessing of the pitch curve from pYIN under PES process.

In Figure 1(B), a segment of *The Love of the Wei River* with mixed playing techniques is taken as an example. The tremolo composed by a series of intensive plucking, vibrato, sliding techniques imitate the vocal style of Qin Opera by controlling the intensity and pitch. Figure 1(C) presents a novel time-frequency representation, and the pitch curves red points were extracted from this representation. The gaps in CTPM through time correspond to the attack-transient intervals in which the signals don't exhibit the periodicity. In the traditional annotation process, the pitch throughout time could be segmented based on the detected boundaries. The start and the end of a tremolo note are often derived from the musician's professional experience and the reference score that reflect a certain subjectivity, thus we manually implement the note segmentation under tremolo and don't delve into its automation

in this paper. As shown in Figure 1(D), the cubic interpolation was used to fill the pitch intervals within the tremolo notes. This step effectively removes abnormal pitch fluctuations at plucks and helps the future analysis of pitch shift techniques under tremolo.

In this paper, we investigate the pitch annotation process and algorithm performance on the string displacement of expressive pipa music by examining boundary detection in terms of recall, F-score and resolution and pitch estimation in terms of Intersection over Union (IoU), pitch deviation and running time. In section II, we briefly review the related works of commonly used onset detector and pitch estimators for annotation. The proposed fully time-domain methods on boundary and pitch were elaborated in Section III. The experimental results and related discussion were assigned in Section IV. Section V concludes this paper and enumerates the improvement and promising research points.

## 2. Related Works

The monophonic transcription consists of the unvoiced/voice detection (UVD) (or called voice activity detection (VAD) [10]) and pitch estimation steps. The annotation of a new particular instrument is generally implemented by completely explainable unsupervised algorithms that can be categorized into parametric, non-parametric and time-domain approaches, where the former two hypothesize the local stationarity of the short-time signal. A typical parametric approach, non-linear least square (NLS [11]), decomposes the signal into the linear combination of the Fourier bases adhering to the property of the stationary waves and harmonic structure. Shi proposed the Bayesian non-linear least square (BNLS [12]) by modeling the order prior and temporal continuity on pitch and voice activity. Most of the non-parametric approaches like the autocorrelation function (ACF [13]), capture the period of speech and music signal and neglect the timbral characteristics. A state-of-the-art pYIN [14] addresses not only the impact from local strength changes, but also temporal continuity on pitch and voice activity. However, the UVD/VAD of the pYIN and BNLS based on the framewise signal reconstruction quality may not meet the complex music structure [2].

As a task to identify the sudden change of signal, the boundary detection provides another path to limit the note range. Since deleting is much easier than adding, high recall and high resolution are the prerequisite to the manual annotation. The framewise root mean square (RMS) envelope of the high-pass signal, referred as to log-energy in the following of this paper, has been used in intensive plucking like tremolo [15]. It has a high recall and resolution suitable for annotation. SpecFlux [16], SuperFlux [17] and ComplexFlux [18] are another three common algorithms in which the first one has been also applied to the guitar annotation [8] and the latter two partially suppress the impact of vibrato and tremolo using maximum filtering from spectral perspective.

The time-domain pitch estimators aims to capture a sequence of zero-crossings (ZCs) or extrema as different sorts of pitch markers (PMs), the period pitch achieved by the distance of two adjacent PMs naturally disentangles to amplitude envelopes and doesn't require window size selection. Multiple local candidates within a period may lead to difficulty in pitch marker selection. The empirical mode decomposition (EMD [19]), as a generic algorithm using the mean curve of upper and lower envelopes passing through the extrema, has been already applied to the endpoint detection [20] and pitch estimation [21] in speech. Although, it doesn't ensure the periodicity of intrinsic mode function (IMF).

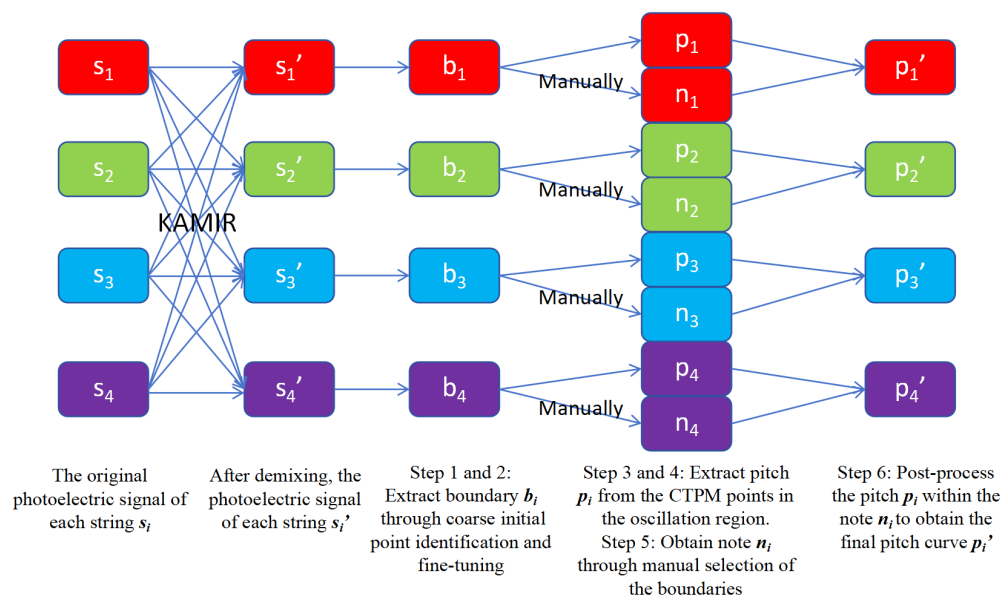
Hess [22] summarized the two categories of time-domain pitch estimators, the first relies on the pre-processing that simplifies the signal structure and reduces the candidate numbers of pitch markers, however a moderate low-pass filter is not easy to select and an arbitrary filter inevitably warp the original signal. DIO algorithm [23] analyzes the PM candidates after low-pass filtering of different cut-off frequencies, this algorithm has a high running speed and comparable performance to the short-time pitch estimators at that time. The second category is based on the thresholding of the raw signal, most of the related works have been appeared before 90s and require a simple waveform structure. Besides, the cascaded approaches first estimate a pitch curve by a robust short-time detector and then captures the PMs with the reference pitch. The granular synthesis [24] using non-positive-to-positive

ZCs and time-scale modification using extrema in the pitch-synchronous overlap add (PSOLA [25]) split the waveform into clips with time-varying lengths referred to the pitch periods obtained by the short-time approaches. In following article, we proposed time-domain approaches on the relatively complex signal without filtering.

### 3. Methodology

The proposed multipitch annotation process based on photoelectric sensors is shown in Figure 2. After removing the string resonance using kernel additive model interference reduction (KAMIR) algorithm [9,26], the monophonic string displacement signals could be processed by the following fully time-domain steps to obtain the APT-required features, i.e. string number, pitch, boundaries, and notes. The first three steps detect the boundaries, the last three process pitch and notes:

1. **Coarse onset detection:** Localize the coarse attack-transient pairs with the arched waveform determined by analyzing the ZC intervals or envelope zero-crossings;
2. **Onset fine-tuning:** Remove the attack-transient pair if the transient doesn't meet the conditions of sudden change to improve the precision. Fine-tune the remaining transients and attacks by searching extrema and inflections;
3. **Vibration signal extraction:** Determine the offset between transient and next attack (the next attack may coincide with offset), extract the the voiced area between transient and offset for pitch analysis;
4. **Pitch marking:** Select an initial pitch marker pair, track forward and backward the pitch markers via the continuous time-period mapping (CTPM);
5. **Manual note segmentation:** Manually select an attack-offset pair as the boundaries of a note;
6. **Pitch interpolation:** Smooth the pitch curve within a note, interpolate the pitch curve clips to fill the gaps within tremolo notes.



**Figure 2.** Multipitch annotation process, the color denotes the string number of corresponding string. Note segmentation is manually implemented in this process.

**Onset Fine-tuning** The sudden change is the key to distinguish the real transient and the impacts of mutual resonance waveform deviating from time axis. In this step, the sudden change is determined by the width ratio of upper and lower envelopes at adjacent extrema points so that a lot of false positives could be eliminated during silence. The correct transient is annotated at the previous extrema before the sudden change. In Figure 3(B), the correct attack is located at the inflection after the detected coarse attack instant produced by finger release. Similarly, the inflection is determined by the soar in

the absolute value of the mean curve slope. The attack-transient interval range approximately between 3.5-150 ms also provides a threshold to remove the false positives from silence areas.

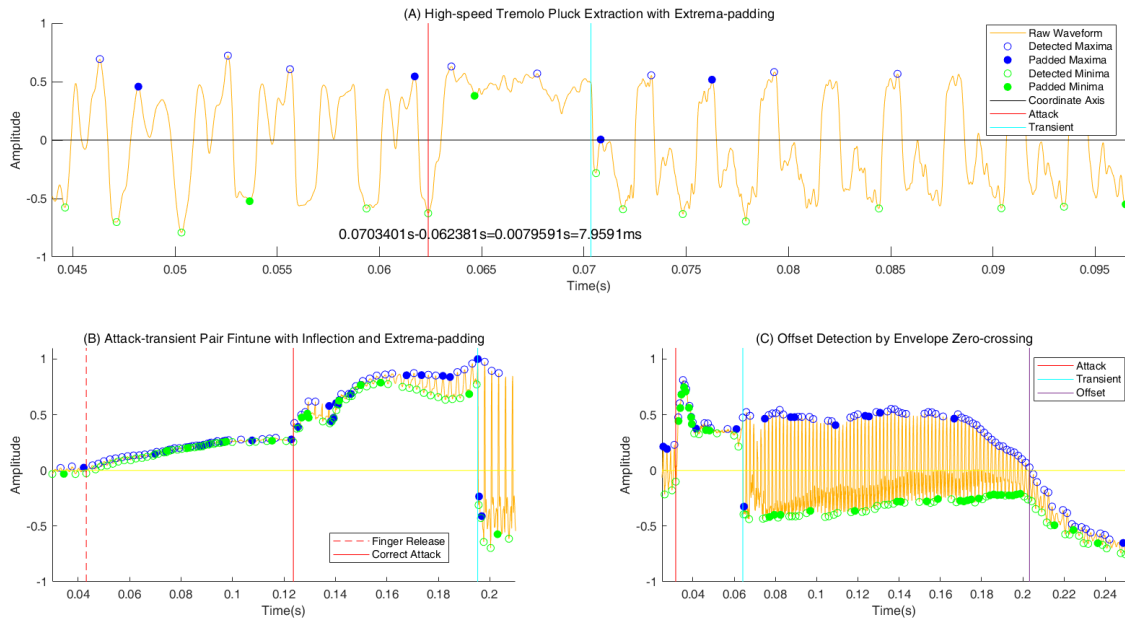


Figure 3. Specific examples for the boundary detection.

**Vibration Signal Area Extraction** As shown in Figure 3(C), the offset of a note is defined at the extrema closest to the instant when the envelope passes through the axis after transient. Therefore, given the boundaries with a sample-level resolution, a clip of signal within transient-offset pair in non-tremolo note and multiple clips in tremolo note count in voiced areas for subsequent pitch estimation.

### 3.1. Pitch Estimation

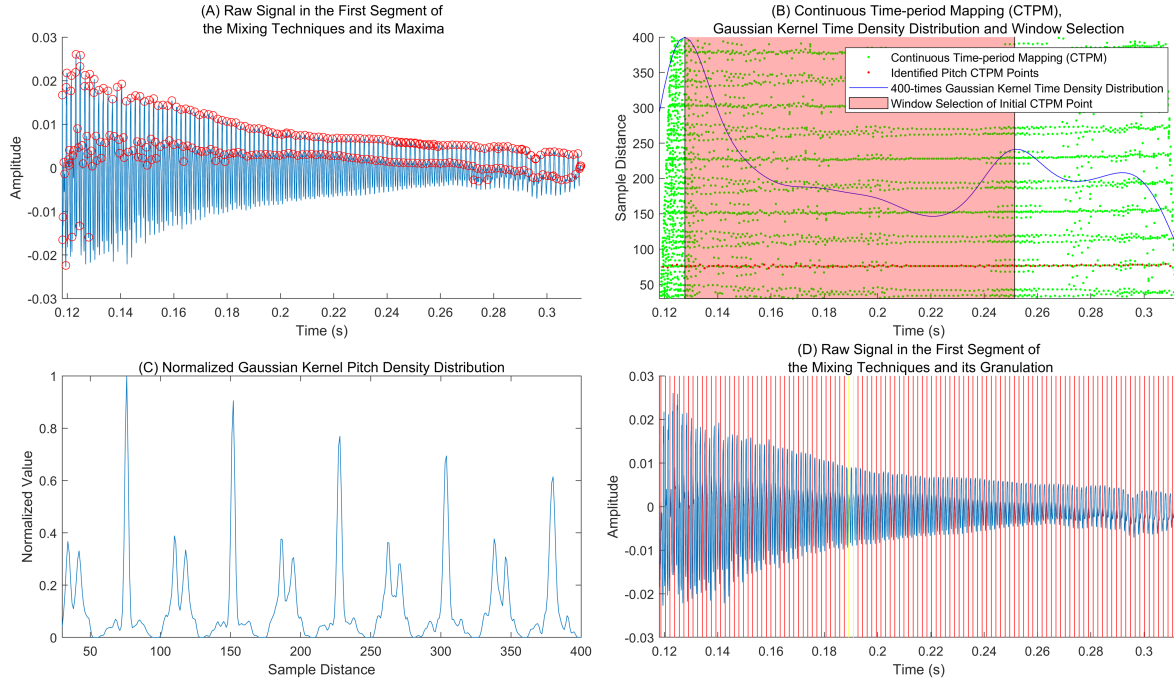
The pitch estimation process is shown in Figure 4. First, we extract zero-crossings or extrema with a minimal distance of 9.09 ms as shown in Figure 4(A). Next, the CTPM and its time-domain density function are calculated to select a window (See Figure 4(B)). This window help to extract the initial pitch points through the pitch density function (See Figure 4(C)). The yellow bars in Figure 4(D) represent the pairs of extrema points corresponding to the initial CTPM points. Starting from these initial point pair, the tracked pitch markers in both directions indicated by red lines in Figure 4(D) also correspond to the red CTPM pitch points shown in Figure 4(B). Some details in the process are described by the follows:

**Continuous Time-Period Mapping (CTPM)** Since the pitch markers are not easy to select if a few of ZC or extrema candidates occur within a period, we propose a novel time-period representation to analyze the pitch structure of signal. The continuous-valued time instant of the  $i^{th}$  pitch marker candidate  $t_i$  is achieved by the linear interpolation for zero-crossings and the parabolic interpolation for extrema. To avoid the excessive shift and envelope asymmetry of signal, we eliminate the mean curves of upper and lower envelopes (the first IMF of EMD) from the split signal. The relevant formulas are described by the follows:

$$p_{i,j} = \frac{1}{t_{i+j} - t_i} \quad T_{i,j} = \frac{t_{i+j} + t_i}{2} \quad (1)$$

$$t_i = \frac{1}{2p_{i,j}} - T_{i,j} \quad t_{i+j} = \frac{1}{2p_{i,j}} + T_{i,j} \quad (2)$$

With  $i, j \geq 1$ , the pitch period  $p_{i,j}$  and its time instant  $T_{i,j}$  of a point in CTPM has the one-to-one correspondence with  $t_i$  and  $t_{i+j}$ , only the distance  $t_{i+j} - t_i$  following the register of particular instrument counts in CTPM candidate points.



**Figure 4.** Specific process for pitch estimation: (A) The first segment of mixed playing technique signal in Fig 1(B) and its maxima; (B) Compute the CTPM, Gaussian kernel density function along the time axis and the pink region used to estimate the initial pitch point; (C) Gaussian kernel density function through the sample distance (pitch) axis of the CTPM points in the pink region; (D) Forward-backward track the pitch marker sequence, the yellow lines represent the extrema pair of the initial CTPM point.

**Pitch Marker Pair Initialization** Known a reference pitch, the pitch marker is usually initialized from the extrema in the entire waveform [27]. Herein, we propose to select the basin of time-domain CTPM marginal distribution as the pitch-sparse window with time-varying length to extract the initial period (See pink area in Figure 4(B)) and the approximate initial instant is defined at the lowest point of the basin. Since CTPM points are continuously valued in both time and period, a Gaussian kernel density estimation method [28] is used to implement the marginal distribution. Assuming the timestamps of  $K$  CTPM points within a certain range are  $T_k$  ( $k = 1, \dots, K$ ), the corresponding time-domain distribution density can be formulated by the follows:

$$P(t) = \frac{1}{K10^{-2}\sqrt{2\pi}} \sum_{k=1}^K \exp\left(-\frac{(t - T_k)^2}{2 \times 10^{-4}}\right) \quad (3)$$

where the standard deviation of the Gaussian kernel is  $10^{-2}$  second. After obtaining the  $P(t)$  distribution, the extrema serve as two ends of the basin. The minimal window length is set to the maximum of 0.1 second and the total length of the audio segment. Once the window is chosen, similarly to equation 3, the period-domain marginal distribution  $P(I)$  of CTPM candidate points can be implemented by equation 4, assuming the variance of the Gaussian kernel as one sample distance and the periods of  $L$  CTPM points within the selected range  $I_l$  ( $l=1, \dots, L$ ):

$$P(I) = \frac{1}{L\sqrt{2\pi}} \sum_{l=1}^L \exp\left(-\frac{(I - I_l)^2}{2}\right) \quad (4)$$

Inspired by the non-parametric pitch estimation methods like ACF [13], we propose to use the first peak greater than  $\lambda \times$  the maximum with  $\lambda \in (0, 1)$  as the approximate period of initial CTPM point so that the corresponding pitch marker pair can be achieved, and the other pitch markers can be inferred before and after this pair using equation 5.

**Pitch Mark Tracking** The continuous reference interval  $I_{n+1}$  is updated by the current interval  $I_n$  and the distance  $t_{i_n \pm j_n} - t_{i_n}$ , and a momentum parameter  $\beta$  is used for the pitch marker tracking:

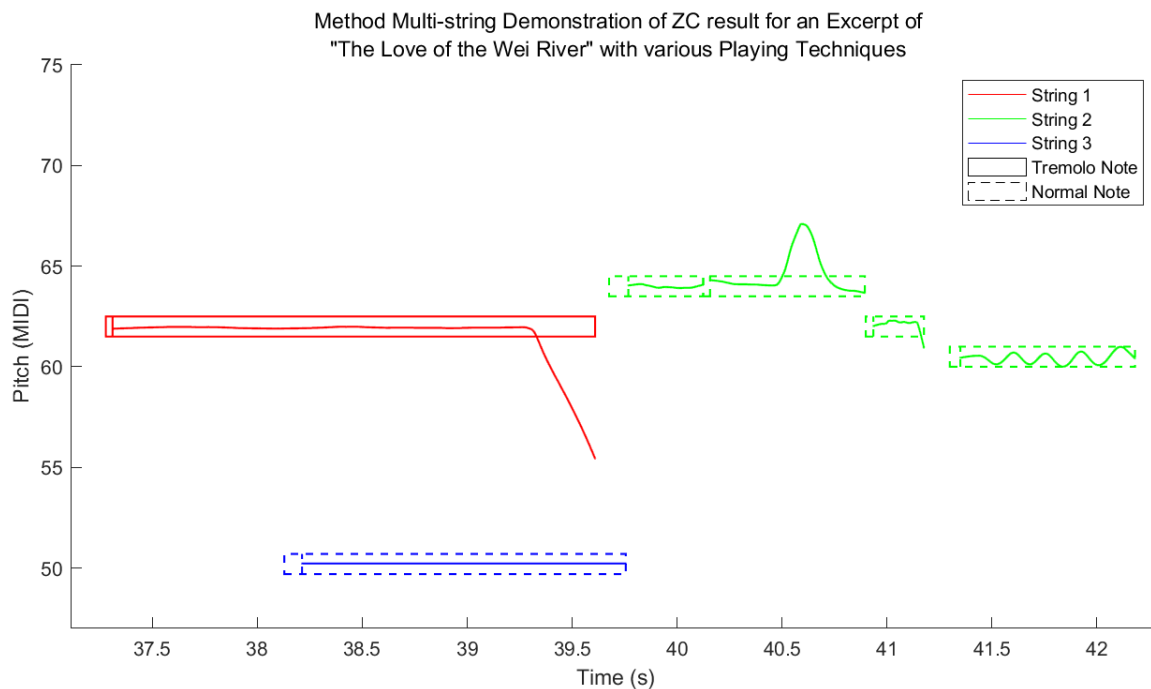
$$I_{n\pm 1} = \beta I_n + (1 - \beta)(t_{i_n \pm j_n} - t_{i_n}) \quad (5)$$

where  $j_n = \arg \min_j (|t_{i_n \pm j} - (t_{i_n} \pm I_n)|)$ ,  $n$  represents the number in the PM sequence, and  $i_{n\pm 1} = i_n \pm j_n$ .

### 3.2. Annotation Process and Processing of the Pitch Curves in Tremolo Notes

Above all, we categorize the processes into the pitch estimation before segmentation (PES) and the segmentation before pitch estimation (SPE). The original pYIN and BNLS attributed to PES process determine the voice activity by the quality of signal reconstruction after pitch estimation. Our proposed approach belongs to the SPE process in which the boundary detection provides a structural prior to constrain the pitch ranges to estimate. The pitch curve throughout the time could be also achieved by setting the voicing probability to 1 in pYIN and BNLS, then assigned by the corrected boundaries into each voiced clip. VAD from original pYIN and BNLS is thus replaced by the correct boundaries and we also name this process directly as PES in this paper.

Subsequently, the attack-offset pairs need to be manually selected through the music score or musicians' experience to implement the note segmentation. The pitch gaps within tremolo notes are finally filled by cubic interpolation for future studies of playing techniques. Notably, the pitch curves need to be smoothed by an outlier-robust method, e.g. robust locally weighted scatterplot smooth (rLoWeSS [29]) algorithm used in this paper, to ensure the pitch continuity after the proposed pitch marking. Figure 5 demonstrates the final APT annotation of another expert of **The Love of the Wei River** including polyphonic notes, tremolo, vibrato, bending and sliding techniques, where the first attack-transient intervals in the tremolo notes like those in the normal notes are not filled.



**Figure 5.** Final multipitch annotation features for an expert of **The Love of the Wei River** with multiple types of playing techniques.

## 4. Experiments

### 4.1. Datasets and Evaluation Metrics

In this section, we examine the transcription performance on the enhanced PipaSet which consists of 4 excerpts of naturally performed famous traditional Chinese music pieces from the collection of pipa grade examination [7], i.e. two low-level pieces **Jasmine Flower** (LvL. 1) and **Nanni Bay** (LvL. 2) and two high-level ones, **The Love of the Wei River** (LvL. 8, new added) and **Ambush from Ten sides** (LvL. 8). The high-level pieces are characterized by longer durations, higher polyphony, and higher degree of technical complexity. Only vibrato and sliding are presented as playing techniques in **Jamine Flower**. The tremolo techniques in **Nanni Bay** and **The Love of the Wei River** are not particularly fast. Vibrato, sliding and tremolo even their mixed ones are widely present in **The Love of the Wei River** and **Ambush from Ten Sides** in which the latter has a large quantity of tremolo plucks with attack-transient intervals shorter than 9.09 ms. The other playing techniques like harmonic, twisting and percussion take a very small proportion. 4 channels of photoelectric signals totaling 984 seconds were recorded with 44100 Hz of sampling rate and 24 bit of depth.

The boundaries are re-annotated following the framework of sample-level precision as outlined in Section 3.1. Since the attacks and transients are only distinguished by the proposed methods in this paper, we merge them together for evaluation with the other algorithms. Moreover, only the initial attack-transient pair of tremolo note is considered in note onset evaluation regardless of the ground-truth or the estimated ones, as all the playing techniques are excluded in a note segment. The accuracy metrics F-score (F), precision (P) and recall (R) are implemented by `mir_eval` [30] with a tolerance of 50 ms ( $\pm 25$  ms). In order to avoid the impact of the note numbers performed on different strings and measure the annotation efficiency, the final F/P/R is weighted by pluck number ratio of each string. A note is identified as a tremolo if at least 2 plucks or 4 peaks present. The corresponding metric, Tremolo Note Recognition (TNR), is shown in Table 2. Given the correct note segments, the results on tremolo pluck localization are examined by the Pluck Detection (PD) metric. Additionally, we select the Intersection over Union (IoU) to evaluate the VAD results of existing pitch algorithms and the proposed boundary detectors. Due to the unavailability of a perfect reference F0, the median of the pitch estimated using the PES/SPE processing method is rounded to the integer MIDI value and compared with the correct MIDI pitch height. Notice that pitch continuity between segments is eliminated in SPE process based method.

Through statistical analysis and parameter scanning, the following parameters were obtained: envelope width ratio (EWR)  $> 0.77$ , sudden slope change (SSC)  $> 2.5$ , momentum  $\beta = 0.8$ , pitch peak ratio  $\lambda = 0.765$ , and rLoWeSS range multiplier of 0.1.

### 4.2. Experimental Results

#### 4.2.1. Boundary Results

According to the results in Table 1, the extrema-padding method achieves 6% and 24% recall improvements respectively to those of log-energy and SpecFlux. The extrema (without padding) and ZC based methods have the same overall F-score, thus 19.25% and 8.25% increases compared to the log-energy and SpecFlux. Although the proposed methods and the log-energy have sufficiently high recalls and resolution, our proposed ones generally outperform the log-energy in F-score.

As shown in Table 2, the extrema-padding method has a 14% TNR recall improvement compared to the log-energy, the extrema (without padding) method shows 21.33% and 33.67% TNR F-score increases compared to the SpecFlux and log-energy methods. Extrema-padding method has a 34% recall and 21.66% F-score improvement compared to log-energy under PD metric. Our proposed detection method outperforms other algorithms in terms of F-score for tremolo cases on all pieces, and the padding method achieves better recall and F-score in **Ambush from Ten Sides**.

**Table 1.** Average note onset detection performance and mean absolute error (MAE) of true positives to ground-truths.

Methods	Jasmine Flower		Nanni Bay		The Love of the Wei River		Ambush from Ten Sides	
	F/P/R (%)	MAE (ms)	F/P/R (%)	MAE (ms)	F/P/R (%)	MAE (ms)	F/P/R (%)	MAE (ms)
SpecFlux	80/91/72	21.9	84/83/86	21.3	30/22/46	19.6	57/60/55	18.3
SuperFlux	70/ <b>96</b> /55	21.7	80/ <b>88</b> /73	21.3	27/27/26	21.4	<b>60</b> /72/52	18.0
ComplexFlux	72/ <b>96</b> /58	21.4	78/ <b>88</b> /70	21.3	25/23/28	21.2	59/76/49	17.7
Log-Energy	80/71/92	2.3	58/42/93	2.9	22/13/ <b>92</b>	3.0	47/43/54	<b>5.6</b>
ZC	88/85/92	<b>1.8</b>	<b>89</b> /87/92	<b>2.4</b>	<b>55</b> / <b>40</b> /90	4.0	52/ <b>81</b> /38	6.5
Extrema w/o padding	<b>90</b> /85/ <b>96</b>	2.4	<b>89</b> /84/94	2.5	54/39/88	<b>2.9</b>	51/69/40	6.8
Extrema w/ padding	83/73/95	2.5	81/70/ <b>95</b>	2.7	42/27/ <b>92</b>	3.7	54/43/ <b>73</b>	7.0

**Table 2.** Tremolo note recognition (TNR), pluck detection (PD) performance and mean absolute error (MAE) of true positives to ground-truths.

Methods	Nanni Bay			The Love of the Wei River			Ambush from Ten Sides		
	TNR (%)	PD (%)	MAE (ms)	TNR (%)	PD (%)	MAE (ms)	TNR (%)	PD(%)	MAE (ms)
SpecFlux	77/62/ <b>100</b>	83/97/73	20.3	57/40/ <b>100</b>	67/99/51	17.2	64/88/50	60/98/43	17.9
SuperFlux	87/77/ <b>100</b>	69/97/53	20.1	67/50/ <b>100</b>	45/ <b>100</b> /29	17.8	55/95/39	49/98/33	17.8
ComplexFlux	91/83/ <b>100</b>	62/97/46	19.5	67/50/ <b>100</b>	46/99/30	17.4	55/90/39	43/98/28	17.7
Log-Energy	45/29/ <b>100</b>	86/96/79	3.2	53/36/ <b>100</b>	68/98/52	5.5	63/68/59	67/96/51	<b>6.9</b>
ZC	<b>100</b> / <b>100</b> / <b>100</b>	98/ <b>100</b> /96	3.0	<b>89</b> / <b>80</b> / <b>100</b>	93/99/87	6.5	69/ <b>100</b> /52	65/ <b>99</b> /48	8.0
Extrema w/o padding	95/91/ <b>100</b>	<b>99</b> / <b>100</b> / <b>99</b>	<b>2.7</b>	<b>89</b> / <b>80</b> / <b>100</b>	<b>98</b> /99/96	<b>4.7</b>	78/97/65	74/98/60	7.5
Extrema w/ padding	56/38/ <b>100</b>	<b>99</b> /98/ <b>99</b>	2.8	62/44/ <b>100</b>	<b>98</b> /97/ <b>100</b>	5.0	<b>82</b> /73/ <b>93</b>	<b>89</b> /95/ <b>85</b>	7.1

Both log-energy and the proposed algorithms correctly estimate samples with a resolution that meets the annotation requirements. However, the latter method has more zero deviation at corrected sample points so that only few boundaries requires manual adjustments. our proposed methods reduces the MAE by approximately 80% compared to the first three short-time methods. Finally, the Table 3 validate the effectiveness of the boundary detection in VAD metric comform to the annotation process of plucked string instruments [2,8].

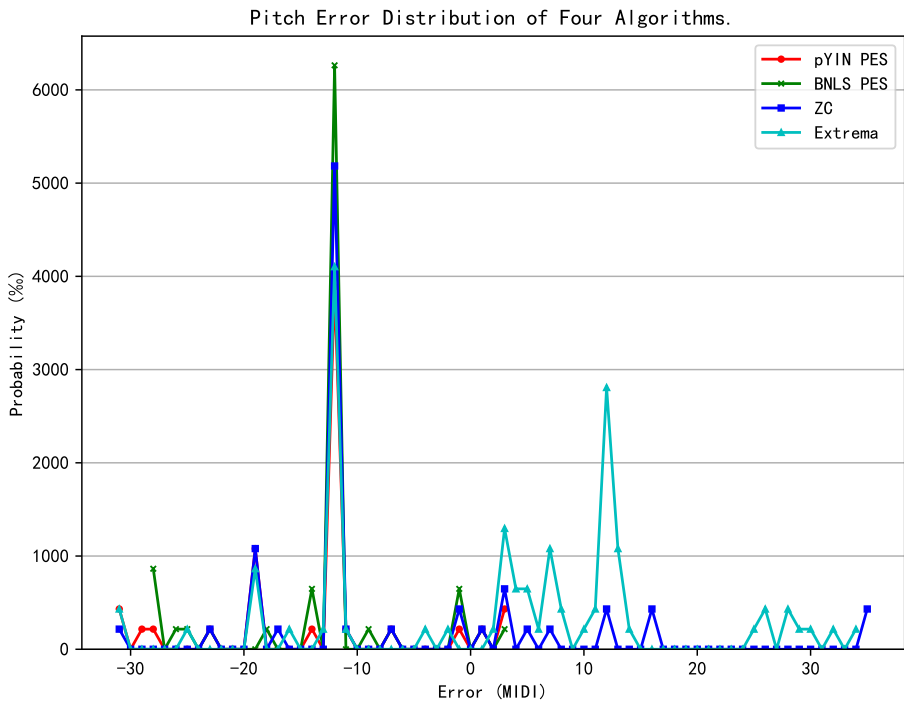
**Table 3.** Average Intersection over Union (IOU) to evaluate voice activity detection and coarse pitch estimation accuracy under PES and SPE processes. The voicing results of the extrema-based approaches w/ and w/o padding are both indicated.

Metrics	Methods			
	pYIN (%)	BNLS (%)	ZC (%)	Extrema (%)
VAD	42.6	34.7	<b>60.1</b>	51.3/50.5
PES	<b>90.9</b>	89	N/A	N/A
SPE	29.2	29.6	<b>87.7</b>	83.9

4.2.2. Pitch Results

According to the results in Table 3, the ZC-based pitch annotation has a slight performance loss compared to the pYIN and BNLS under PES process. Very catastrophic results arise in pYIN and BNLS under SPE process.

Finally, Figure 6 presents a MIDI-level pitch deviation (estimated vs. true) among the proposed pitch estimation algorithms and the two short-time ones. It shows that the errors from the all algorithms primarily concentrate in the sub-harmonic notes corresponding to the octave errors.



**Figure 6.** Global pitch error histogram

#### 4.2.3. Running Speed

All the algorithms proposed in this paper are implemented in Matlab. Even without any code optimization, the method still has overwhelming speed advantages. Running on hardware with an AMD R7-5800H CPU, 32GB RAM and Matlab 2023b with the correct boundaries, the zero-crossing-based pitch estimation demonstrates a 26 and 27 times speed-up to the C++ implemented pYIN<sup>1</sup> in the SPE and PES processes, 108 and 65 times compared to the Matlab version of BNLS<sup>2</sup>.

### 5. Discussion

The results of boundary detection and tremolo analysis depict a significant improvement on the first three excerpts, the recalls in **Ambush from Ten Sides** decrease due to the loss of boundaries caused by the insufficient attack-transient intervals shorter than 9.09 ms produced by high-speed tremolo. The extrema-padding alleviates this issue at the expense of a precision decrease. Therefore, we recommend the extrema-padding onset detector for very intensive tremolo plucking annotation if the tremolo rate exceeds 10 Hz.

For quasi-periodic part of signals, the extrema generally generate more candidates than the ZCs, this results in higher difficulty in initial pitch selection than ZC and an inferior pitch performance referring to Figure 6. Very catastrophic results arise in pYIN and BNLS under SPE process as shown in Table 3, this implies that our proposed method can effectively resist the non-stationarity caused by AM-FM signals in relatively short time, while short-time approaches rely on the signal context and inherent pitch continuity for better results to some extent.

Since the tremolo and pitch shift techniques require theoretically high resolution respectively in time and frequency, the window selection become tough for high-resolution music transcription. The key point of a signal and the CTPM proposed in this paper are appropriate solutions to deal with AM-FM characteristics and the lack of pitch synchronism in EMD-like algorithms, but the noise-sensitive drawback still exists. Owing to the photoelectric sensor which have a direct contact with the vibration source, the external noise is excluded so that this system has the similarity to the guitar pickup that can be applied to a live performance full of complex ambient noise. Meanwhile according to the sound synthesis theory of pipa instrument[31], the bass audio recordings may produce a salient high-order harmonic components comparing to the F0 due to the small body size of pipa, the string vibration not modulated by the body resonance reflects waveform structure with fewer oscillation within a period.

### 6. Conclusion and Future work

This paper proposes a high-resolution transcription and annotation algorithm for pipa photoelectric signals, based on zero-crossing and extremum points in the full time domain. The method leverages the morphological advantages of photoelectric sensor signals over audio and electromagnetic signals during plucking. Additionally, a novel time-period representation method, CTPM, is introduced for waveform structure visualization and pitch recognition tasks. After benchmarking the performance of 6 short-time onset detection and pitch estimation algorithms in complex playing technique, it is evident that the proposed zero-crossing and extremum-based boundary detection method achieves superior F-scores in note onset recognition while maintaining a high recall rate. In high-speed tremolo scenarios, the extrema boundary detection method with padding-point improves the recall rate significantly. For the pitch estimation of vibration signal, our proposed method has ultimately achieved a relatively high level of performance and greater efficiency. Combining with the post-processing, the AMT task on mixed playing techniques could be addressed by our proposed annotation process.

<sup>1</sup> pYIN code available: <https://code.soundsoftware.ac.uk/projects/pyin>

<sup>2</sup> BNLS code available: <https://github.com/LimingShi/Bayesian-Pitch-Tracking-Using-Harmonic-model>

In the future, we will focus on improving the algorithm in terms of accuracy and resolution, particularly in the specific scenarios such as high-speed tremolo. The dataset will also be scaled up to further validate the reliability of the algorithm under a funding afterwards. The proposed CTPM, as a novel time-frequency representation, holds a potential for more exploration in signals from other sensors, like the audio and electromagnetic pickup. Additionally, the transcribed pitch curves can be used for pitch shift technique analysis using similar morphological methods. Finally, the single-track multipitch estimation remain challenging and promising as the future direction.

**Author Contributions:** Conceptualization, Y. W. and Q.W.; methodology, Y. W. and Q.W.; software, Y. W., X. L., Y. Z.; writing—original draft preparation, Y. W.; writing—review and editing, X. L. and Q. W.; visualization, X. L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** The study does not include ethical issues.

**Data Availability Statement:** The data presented in this study are openly available in [https://github.com/veneno1213822/CTPM\\_DATA](https://github.com/veneno1213822/CTPM_DATA).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Palmer, C.; Hutchins, S. What Is Musical Prosody? *Psychology of Learning and Motivation* **2016**, *46*, 245–278.
- Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. *Proc. TENOR*, 2015, p. 23–31.
- Ting-Wei, S.; Yuan-Ping, C.; Li, S.; Yi-Hsuan, Y. Tent: Technique-embedded note tracking for real-world guitar solo recordings. *Transactions of the International Society for Music Information Retrieval* **2019**, *2*, 15–28.
- Yu, Z.; Ziya, Z.; Xiaobing, L.; Feng, Y.; Sun., M. CCOM-HuQin: An Annotated Multimodal Chinese Fiddle Performance Dataset. *Transactions of the International Society for Music Information Retrieval* **2023**.
- Bochen, L.; Xinzhaoh, L.; Karthik, D.; Zhiyao, D.; Gaurav, S. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Trans. Multimedia* **2019**, *21*, 522–535.
- Helena, C.; Emilia, G. Voice assignment in vocal quartets using deep learning models based on pitch salience. *Transactions of the International Society for Music Information Retrieval* **2022**, *5*, 99–112.
- Pipa Professional Committee of Shanghai Musicians Association. *The Collection for Pipa Grade Examination of China*; Shanghai Music Publishing House, 2012.
- Xi, Q.; Bittner, R.; Pauwels, J.; Ye, X.; Bello, J. GuitarSet: A dataset for guitar transcription. 19th International Society for Music Information Retrieval Conference (ISMIR), 2018, p. 453–460.
- Wang, Y.; Jing, Y.; Wei, W.; Cazau, D.; Adam, O.; Wang, Q. PipaSet and TEAS: A Multimodal Dataset and Annotation Platform for Automatic Music Transcription and Expressive Analysis dedicated to Chinese Traditional Plucked String Instrument Pipa. *IEEE ACCESS* **2022**, *10*, 113850–113864.
- Drugman, T.; Stylianou, Y.; Kida, Y.; Akamine, M. Voice Activity Detection: Merging Source and Filter-based Information. *IEEE Signal Processing Letters* **2016**, *23*, 252–256.
- Quinn, B.G.; Thomson, P.J. Estimating the frequency of a periodic function. *Biometrika* **1991**, *78*, 65–74.
- Shi, L.; Nielsen, J.K.; Jensen, J.R.; others. Robust Bayesian Pitch Tracking Based on the Harmonic Model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2019**, *27*, 1737–1751. Code available: <https://github.com/LimingShi/Bayesian-Pitch-Tracking-Using-Harmonic-model>.
- Shimamura, T.; Kobayashi, H. Weighted autocorrelation for pitch extraction of noisy speech. *IEEE transactions on speech and audio processing* **2001**, *9*, 727–730.
- Mauch, M.; Dixon, S. pYIN: A Fundamental Frequency Estimator using Probabilistic Threshold Distributions. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014. Code v1.1.1 available: <https://code.soundsoftware.ac.uk/projects/pyin/files>.
- Freire, S.; Nézio, L. Study of the tremolo technique on the acoustic guitar: Experimental setup and preliminary results on regularity. *Int. Conf. Sound and Music Computing*, 2013, pp. 329–334.
- Bello, J.P.; Daudet, L.; Abdallah, S.; Duxbury, C.; Davies, M.; Sandler, M. A tutorial on onset detection in music signals. *IEEE Transactions on speech and audio processing* **2005**, *13*, 1035–1047.

17. Bock, S.; Widmer, G. Maximum Filter Vibrato Suppression for Onset Detection. *International Conference on Digital Audio Effects (DAFx)*, 2013.
18. Bock, S.; Widmer, G. Local Group Delay based Vibrato and Tremolo Suppression for Onset Detection. *14th International Society for Music Information Retrieval Conference (ISMIR)*, 2013.
19. Nord, H.; Zheng, S.; Steven, L.; Manli, W.; Hsing, S.; Quanan, Z.; Nai-Chyuan, Y.; Chi Chao, T.; Henry, L. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 1998, Vol. 454, pp. 903–995.
20. Manman, L.; Hongwu, Y.; Weitong, G.; Dong, P.; Hongying, S. Endpoint Detection of Noisy Speech using Empirical Mode Decomposition. *International Journal of Digital Content Technology and its Application (JDCTA)* **2012**, 6, 196–203.
21. Huang, H.; Pan, J. Speech pitch determination based on Hilbert-Huang transform. *Signal Processing* **2006**, 86, 792–803.
22. Hess, W., Pitch and Voicing Determination of Speech with an Extension Toward Music Signals. In *Springer Handbook of Speech Processing*; 2008. Available online: <https://api.semanticscholar.org/CorpusID:16346607>.
23. Morise, M.; Kawahara, H.; Katayose, H. Fast and Reliable F0 Estimation Method Based on the Period Extraction of Vocal Fold Vibration of Singing Voice and Speech. *Physics* **2009**.
24. Wishart, T. *Audible Design: A Plain and Easy Introduction to Practical Sound Composition*; Orpheus books, 1994.
25. Moulines, E.; Charpentier, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.* **1990**, 9, 453–467.
26. Pratzlich, T.; Bittner, R.M.; Liutkus, A.; Müller, M. Kernel additive modeling for interference reduction in multi-channel music recordings. *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2015, p. 584–588. Code available: <https://members.loria.fr/ALiutkus/kamir/>.
27. Cheng-Yuan, L.; Jyh-Shing, R.J. A two-phase pitch marking method for TD-PSOLA synthesis. *INTER-SPEECH*, 2004.
28. Sen, B. Introduction to the Nonparametric statistics, n.d. Columbia University. Available online: <https://www.stat.columbia.edu/~bodhi/Talks/Intro&NP-Stat.pdf>.
29. Cleveland, W.S. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* **1978**, 74, 829–833.
30. Raffel, C.; McFee, B.; Humphrey, E.J.; others. mir\_eval: A transparent implementation of common mir metrics. *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*, 2014, pp. 367–372.
31. Chen, Y.H.; Huang, C.F. Sound Synthesis of the Pipa Based on Computed Timbre Analysis and Physical Modeling. *IEEE Journal of Selected Topics in Signal Processing* **2011**, pp. 1070–1079.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.