Advances in Computer Science Research, volume 65

3rd International Conference on Communications, Information Management and Network Security (CIMNS 2018)

A Frame-Based Extended MIDI-Pitch Detection Algorithm Based on Variable Sampling Technology

Yin Feng* and Mingtai Lin

Cognitive Science Dept., School of Information Science and Technology, Xiamen University, P.R. China

*Corresponding author

***Abstract*—This paper proposes a framed-based extended MIDI-pitch detection algorithm based on variable sampling technology. Symbols *P*--, *P*-, *P*+ and *P*++ are used to present four different logarithmic frequencies around the MIDI pitch *P* (integer) within 100 cents interval and called as extended MIDI-pitch to enhance the expression resolution of MIDI pitch of a signal frame. Experiment shows that the algorithm has a good raw pitch accuracy rate and less computation complexity o(nlog$_2$n).**

Keywords—extended MIDI pitch; singing transcription; singing signal processing

I. INTRODUCTION

Since the 1970s, some scholars have been engaged in research on the singing transcription system and have made certain achievements. Singing transcription can be performed at different levels: low-level description (fundamental frequency and energy of signal frame) and higher structure levels (note segmentation, tonality analysis and estimation of note pitch)[1].

The extraction method of fundamental frequency of the singing signals can be divided into the time domain method [2]-[4], frequency domain method [5]-[12] and time-frequency combination method [13]-[15]. The present technology regarding extraction of the fundamental frequency of signal frame can basically satisfy the high-level analysis demand. The YIN algorithm [2] put forward by Chevegigne makes improvements in the cumulative average normalized function used in the autocorrelation function, and it is good in accuracy but not quite fast in the arithmetic speed. The computation complexity of YIN algorithm is o(n^2). The remainder of this paper is organized as follows: A framed-based pitch (frequency) detection algorithm with less computation complexity o(nlog$_2$n) is presented in section II. The experimental result and discussion will be shown in section III. The conclusion will be stated in section IV.

II. FRAME-BASED EXTENDED MIDI-PITCH DETECTION ALGORITHM

A new pitch (i.e. fundamental frequency) detection algorithm with computation complexity o(nlog$_2$n) for singing waveform is proposed in this paper. Our acoustics representation is based on the ubiquitous and well-known spectrogram, which converts the singer's singing waveform into a distribution of energy over time and frequency. The original singing recordings are combined into a single (mono) channel with 44 kHz sampling rate. We apply a fast Fourier transform (FFT), using *N*=4096 point transforms, an *N*–point Hamming window, a 2048 point overlap of neighboring windows.

Let $f_{t,j}$, $E(f_{t,j})$ and $\{ f_{t,j} \}$ denote respectively the frequency, the energy of $f_{t,j}$ with respect to FFT index j and the set consisting of the frequency of FFT in frame t. Then there may exist the $f_{t,j}^m \in \{ f_{t,j} \}$ in frame t such that $f_{t,j}^m = mf_{t,j} \Delta f_{t,j}^m$, where $f_{t,j}^m$ is the mth harmonic on the frequency $f_{t,j}$, $\Delta f_{t,j}^m$ is some value in the number interval $[2^{-\frac{1}{24}}, 2^{\frac{1}{24}})$. $\Delta f_{t,j}^m$ is introduced because using FFT to estimate the frequency is restricted by the sampling rate in the accuracy. That is, regard the frequency of the frame t estimated through FFT in the defining section $[2^{-\frac{1}{24}} mf_{t,j}, 2^{\frac{1}{24}} mf_{t,j})$ as the $f_{t,j}^m$ of $f_{t,j}$. Then the sum of energy of frequencies of $f_{t,j}$ and its harmonics can be achieved by the formula $\sum_{m=1}^{M} E(f_{t,j}^m)$, where M is the number of $f_{t,j}$'s harmonics $(f_{t,j} = f_{t,j}^1)$. Therefore, the equation (1) can be used to estimate the fundamental frequency $f0_t$ of the frame t.

$$f0_t = \underset{f_{t,i} \in \{f_{t,i}\}}{arg\,max} \sum_{m=1}^{M} E\left(f_{t,i}^m\right) \tag{1}$$

Wei-Ho Tsai and Hsin-Chieh Lee [11] also used the similar method, namely, using the equation (1) to estimate the fundamental frequency. Actually, the estimation result of the equation is not reliable. Sometimes the 2nd ~ 5th harmonics will be misjudged as the fundamental frequency or conversely; that is, the 2nd ~ 5th harmonics of the estimation result is the correct fundamental frequency. Besides, the pitch accuracy is not enough to achieve the identification standard that can be accepted by the singing transcription expert. We will describe details of reducing the fundamental frequency misjudgment rate and improving the estimation accuracy of the pitch below.

A. $F0_t$ Correction Model

The $f0_t$ estimated according to the equation (1) is further corrected as $F0_t$ based on six situations in Table 1. Take M=6 during the experiment.

The fundamental frequency detection algorithm described in the Table 1 is the temporary experience equation obtained after the experienced singing transcription expert invited by the author observes signal frame fundamental frequency estimates in the singing signal area and changes in the energy ratio of the 2nd ~ 6th harmonics to the fundamental frequency and compares them with the subjective pitch hearing of the singing transcription expert through experiment. With the growing of the experiment data, the equation may also change.

TABLE I. F0 CORRECTION MODEL

$F0_t$	Experience equation
$5f0_t$	Energy ratio of the 5th harmonics of the current frame to the fundamental frequency ≥ threshold value 10
$4f0_t$	Energy ratio of the 4th harmonics, 3rd harmonics and 5th harmonics of the current frame to the fundamental frame ≥ threshold value 10
$3f0_t$	Situation 1: energy ratio of the 3rd harmonics of the current frame to the fundamental frequency ≥ threshold value 3.5 and: 1) energy ratio of the 3rd harmonics of the current frame to the 2nd harmonics or the 4th harmonics ≥ threshold value 4; or 2) energy ratio of the 5th harmonics of the current frame to the 4th harmonics ≥ threshold value 4
	Situation 2: energy ratio of the 3rd harmonics of the current frame to the fundamental frequency ≥ threshold value 3.5; and energy ratio of the 6th harmonics to the fundamental frequency ≥ threshold value 4; and energy ratio of the 4th harmonics to the 3rd harmonics ≤ threshold value 10; and energy ratio of the 4th harmonics to the 5th harmonics ≤ threshold value 2;
	Situation 3: energy ratio of the 3rd harmonics of the current frame to the fundamental frequency ≥ threshold value 9; and energy ratio of the 4th harmonics to the 3rd harmonics ≤ threshold value 10; and energy ratio of the 4th harmonics to the 5th harmonics ≤ threshold value 10;
$2f0_t$	Situation 1: threshold value 5 ≤ energy ratio of the 4th harmonics of the current frame to the fundamental frequency < threshold value 10; and energy ratio of the 4th harmonics to the 5th harmonics ≥ threshold value 1.5;
	Situation 2: threshold value 6 < energy ratio of the 2nd harmonics of the current frame to the fundamental frequency < threshold value 20; and energy ratio of the 2nd harmonics to the 3rd harmonics and 5th harmonics ≥ threshold value 3;
	Situation 3: energy ratio of the 3rd harmonics of the current frame to the fundamental frequency ≥ threshold value 20
$1.5f0_t$	Threshold value 3.5 ≤ energy ratio of the 3rd harmonics of the current frame to the fundamental frequency < threshold value 9; and energy ratio of the 3rd harmonics to the 2nd harmonics≥ threshold value 3.5; and energy ratio of the 6th harmonics to the fundamental frequency ≤ threshold value 4;
$f0_t$	Otherwise

B. Signal Frame Extended MIDI Pitch Estimate Based on Variable Sampling Technology

We adopt a variable sampling technology to estimate the frame's MIDI pitch. On the one hand, it can improve the estimation accuracy (see the extended MIDI pitch in Definition 5 later for details); on the other hand, it can further reduce the probability of the harmonics misjudged as the fundamental frequency.

Judith C. Brown [7] shows us a result that changing the time duration of a signal frame can adjust the value estimated on the frequency of the signals frame through FFT. Similarly, when the sampling rate changes, the time duration of a signal frame (with same samples) changes, and the frequency of the frame estimated by FFT program will also change. For example, sampling rate of the singing signal is reduced from 44 kHz to 22 kHz, the value estimated on the fundamental frequency of the same signal frame (with same samples) through FFT is two times of the original one (namely, the pitch estimate will be improved by one octave). Generally, if the sampling rate of the singing signal is reduced by 1/K times of the original one, the estimated result of the fundamental frequency F0 of the signal frame based on Table 1 will be K times of the original one. We hope to select the appropriate K value to adjust the results of the MIDI pitch of singing signals estimated based on Table 1 and equation (1) successively. According to the equal temperament, take K=1, $\sqrt[48]{2}$, $\sqrt[24]{2}$, $\sqrt[16]{2}$, and $\sqrt[12]{2}$ respectively and reduce the singing signal with 44 kHz sampling rate by 1/K times, and the results of the MIDI pitch of the singing signal estimated based on equation (1), Table 1 and question (2) shown as below should change within the scope of one semitone intervals.

$$\text{MIDI} = \left\lfloor 12log_2\left(\frac{F0}{440}\right) + 69.5 \right\rfloor \quad (2)$$

Therefore, the concept of estimated pitch with variable sampling rate can be introduced as follows.

Definition 1 (Estimated pitch with variable sampling rate): Assume HS_1 is the singing signal with 44 kHz sampling rate. Reduce HS_1 by 1/K times based on K=1, $\sqrt[48]{2}$, $\sqrt[24]{2}$, $\sqrt[16]{2}$, and $\sqrt[12]{2}$, and in proper order. Then the new singing signals HS_2, HS_3, HS_4 and HS_5 are obtained and their sampling rates are as stated in Table II. Assume $p_1(t)$, $p_2(t)$, $p_3(t)$, $p_4(t)$ and $p_5(t)$ are respectively the MIDI pitch of the No. t frame of singing signals HS_1, HS_2, HS_3, HS_4 and HS_5 based on (1), (2) and Table 1. The frame size used during the estimation is 4,096 points with 2,048 overlapped points by the neighboring frames. The 5-tuple < $p_1(t)$, $p_2(t)$, $p_3(t)$, $p_4(t)$, $p_5(t)$ > is called the estimated pitch with variable sampling rate of the No. t frame of the signal HS_1.

TABLE II. SAMPLING RATES OF HS1~HS5

Singing signal with variable sampling rate	HS_1	HS_2	HS_3	HS_4	HS_5
***K* value**	1	$\sqrt[48]{2}$	$\sqrt[24]{2}$	$\sqrt[16]{2}$	$\sqrt[12]{2}$
Sampling rate(Hz)	44,000	43,369	42,747	42,135	41,530

What should be illustrated is that the estimated (frequency) pitch of the signal frame in Definition 3 will change with the change in the sampling rate. Moreover, whether the pitch of the signal frame estimated based on equations (1), (2) and Table 1 can be used to estimate the pitch of a note is still related to whether the selection of the sampling rate and frame size is appropriate. It is put forward in the constant-Q method by Judith C. Brown [7]-[8] that the sampling rate maintains certain relationship with the frequency band of the signal, which can make the estimated value of the fundamental frequency at the center of the hearing pitch frequency; otherwise, there will be certain discrepancies between the estimated pitch of the signal frame and the hearing pitch in the corresponding signal area. Therefore, we might as well stipulate the range and regard the estimated pitch of signal frames out of the stipulated range as the invalid frame of the estimated pitch. The invalid frame of the estimated pitch can be used as the basis to judge the aspirated sound area and breath area in the identification of the boundary of notes and the pitch estimation, but not as the reference value of the pitch estimation.

Definition 2 (valid estimated range, valid frame and invalid frame of the estimated pitch): Assume the MIDI pitch of the No. t frame of the singing signals with one of sampling rate described in Table 2 is the integer P_t which is estimated through the equations (1), (2) and Table 1. The MIDI integer section [P_1, P_2] is called the valid estimated range of the signal frame. If $P_1 \le P_t \le P_2$, the No. t frame of the singing signal is called the valid

frame of the estimated pitch; otherwise (namely, when $P_t < P_1$ or $P_t > P_2$), the No. t frame of the singing signal is called the invalid frame of the estimated pitch.

In consideration of the error possibly occurring during the rounding operation based on the equation (2), under the five sampling rates as stated in Definition 3, the results of the MIDI pitch of the singing estimated through the equations (1), (2) and Table 1 should change within the scope of no more than two semitone intervals (also 200 cents) at most. During the experiment, we take the integer section [41, 76] of the MIDI pitch as the valid estimated range of the signal frame; that is, the signal frame when the MIDI pitch name changes within the scope between F2 to E5 (three octaves) is regarded as the valid frame of the estimated pitch, or it is the invalid frame of the estimated pitch, which is basically proved by the experimental data. Of course, some situations of the harmonics being misjudged as the fundamental frequency still exist in the actual estimated result. Actually, such misjudgment also exists in the YIN algorithm and other autocorrelation-based fundamental frequency algorithms similarly, while the advantage of the estimated fundamental frequency with variable sampling rate lies in that the subtle reduction can also make imperceptible changes in the corresponding signal frame area. Through the intercomparison of five possible fundamental frequency estimated results of the signal frame under 5 sampling rates, the probability of harmonics being misjudged as the fundamental frequency can be eliminated or reduced, which makes the estimated fundamental frequency with variable sampling rate have much lower harmonics misjudgment rate than the existing YIN algorithm or other FFT-based fundamental frequency algorithms. To eliminate the probability of harmonics being misjudged as the fundamental frequency, how to realize the replacement of the harmonics pitch by the fundamental frequency pitch through seeking the approximation relation (the error is within the scope of two semitone intervals or 200 cents) between the fundamental frequency and harmonics of the signal frame under five different sampling rates is described in the following Definition 3 and Algorithm 1.

Definition 3 (harmonics relation of estimated pitch with variable sampling rate): Assume $< p_1(t), p_2(t), p_3(t), p_4(t), p_5(t) >$ as the estimated pitch with variable sampling rate of the No. t frame of the singing signal HS_1. $[P_1, P_2]$ is the valid estimated range of the signal frame. Make $p_r(t)=\min\{ p_1(t), p_2(t), p_3(t), p_4(t), p_5(t)\}$. If $P_1 \le p_r(t) \le P_2$ and $p_s(t)$, $1 \le s(\neq r) \le 5$ making one of the following 1) ~ 3) effective exists, we say the harmonics relation exists between $p_s(t)$ and $p_r(t)$ (or $p_r(t)\pm1$ or $p_r(t)\pm2$).

1) $10 \le p_s(t)-p_r(t) \le 14$, namely, $p_s(t)$ is one octave higher than $p_r(t)$ (or $p_r(t)\pm1$ or $p_r(t)\pm2$);

2) $17 \le p_s(t)-p_r(t) \le 21$, namely, $p_s(t)$ is one octave and perfect fifth higher than $p_r(t)$ (or $p_r(t)\pm1$ or $p_r(t)\pm2$);

3) $22 \le p_s(t)-p_r(t) \le 26$, namely, $p_s(t)$ is two octaves higher than $p_r(t)$ (or $p_r(t)\pm1$ or $p_r(t)\pm2$);

Note: $p_r(t)\pm1$ represents the MIDI pitch semitone interval higher or lower than $p_r(t)$; $p_r(t)\pm2$ represents the MIDI pitch one whole-tone interval higher or lower than $p_r(t)$.

Algorithm 1: Harmonics replacement of the signal frame

Input: the estimated pitch sequence with variable sampling rate of the singing signal HS_1: $<p_1(0), p_2(0), p_3(0), p_4(0), p_5(0)>$, $<p_1(1), p_2(1), p_3(1), p_4(1), p_5(1)>$,…, $<p_1(N), p_2(N), p_3(N), p_4(N), p_5(N)>$. The valid estimated range of the signal frame is $[P_1, P_2]$. Among which, P_1 and P_2 represent the MIDI pitch (integer) of which the pitch names are F2 and E5.

Output: 5-tuple sequence:

$<p_{r_1}(0), p_{r_2}(0), p_{r_3}(0), p_{r_4}(0), p_{r_5}(0)>$, $<p_{r_1}(1), p_{r_2}(1), p_{r_3}(1), p_{r_4}(1), p_{r_5}(1)>$, …, $<p_{r_1}(N), p_{r_2}(N), p_{r_3}(N), p_{r_4}(N), p_{r_5}(N)>$

Algorithm description:

Step 1: *For* $t = 0, \ldots, N$ *do*

Classify the five MIDI pitch with variable sampling rate of the No. t frame, to make the interval of any two pitch with variable sampling rate of the same kind within one whole-tone interval;

If four of the five MIDI pitch with variable sampling rate of the No. t frame fall into the same class,

then make the other MIDI pitch with variable sampling rate not falling into this class ±12 or ±19 or ±24, to make these five fall into the same class;

Step 2: *For* $t = 1, \ldots, N$ *do*

If at least two of the five MIDI pitch with variable sampling rate of the No. t frame fall into the same class and they and all the MIDI pitch with variable sampling rate of the No. t-1 frame belong to the same class,

then make the other MIDI pitch with variable sampling rate not falling into this class ±12 or ±19 or ±24, to make these five fall into the same class;

Step 3: *For* $t = 0, \ldots, N$ *do*

If three of the five MIDI pitch with variable sampling rate of the No. t frame fall into the same class, *then* make other MIDI pitch with variable sampling rate not falling into this class ±12 or ±19 or ±24, to make these five fall into the same class;

Step 4: *For* $t = 0, \ldots, N$-1 *do*

If only two of the five MIDI pitch with variable sampling rate of the No. t frame fall into the same class and they and all the MIDI pitch with variable sampling rate of the No. t+1 frame belong to the same class,

then make other MIDI pitch with variable sampling rate not falling into this class ±12 or ±19 or ±24, to make these five fall into the same class;

Step 5: *For t* = 0 ,…, *N do*

Output all five MIDI pitch with variable sampling rate of the No. *t* frame after adjusted

$<p_{r_1}(t),p_{r_2}(t),p_{r_3}(t),p_{r_4}(t),p_{r_5}(t)>$

Definition 4 (harmonics replacement of estimated pitch with variable sampling rate): Assume the input of Algorithm 1 is the estimated pitch sequence with variable sampling rate of the signal frame of the singing signal HS_1. Call the No. *t* 5-tuple of the output sequence in Algorithm 1 $<p_{r_1}(t), p_{r_2}(t), p_{r_3}(t), p_{r_4}(t), p_{r_5}(t)>$ the harmonics replacement of the estimated pitch with variable sampling rate of the No. *t* frame of the singing signal HS_1.

In Algorithm 1, the MIDI pitch with variable sampling rate of the No. *t* frame is replaced with the most appropriate pitch with variable sampling rate satisfying the harmonics (the 2nd ~ 4th harmonics, namely, MIDI pitch ±12 or ±19 or ±24) relation with the No. *t* frame or No. *t*-1 frame or No. *t*+1 frame as stated in Definition 5, which can reduce the misjudgment rate of the valid frame of the estimated pitch as stated in definition 4.

The following discussions are involved in the more accurate estimation of the pitch of the signal frame. The symbols *P*--, *P*-, *P*+ and *P*++ shown in Figure 1 stand for four different pitch points within 100 cents (i.e. semitone) interval around the MIDI pitch *P* identified by a singing transcription expert based on the hearing. Assume the left and right relationship of the position of these four pitch points in the figure represents the subjective hearing judgment on the pitch of these four singing notes in a section of singing by the experienced singing transcription expert. Namely, the pitch of the left notes is low and the right is high. Five logarithmic number axes below these pitch points (logarithmic frequencies) represent the pitch estimation models for these four notes recorded according to the five different sampling rates as stated before and estimated based on the equations (1), (2) and Table 1 respectively. Their output estimated results can only be the MIDI pitch (integer), namely, *P*-1, *P*, *P*+1 or *P*+2.

Definition 5 (extended MIDI pitch): Assume the set N_{MIDI} is an integer set constituted by all the MIDI pitch. Construct the set $N^{+}_{\mathrm{MIDI}}=\{P++|P \in N_{\mathrm{MIDI}}\}\cup\{P+|P \in N_{\mathrm{MIDI}}\}\cup\{P-|P \in N_{\mathrm{MIDI}}\}\cup\{P--|P\in N_{\mathrm{MIDI}}\}$. Where, *P*--, *P*-, *P*+, *P*++ stand for 4 different pitch points within 100 cents (i.e. semitone) interval around the MIDI pitch *P*. *P*-- is 25 cents lower than *P*-, *P*- is 25 cents lower than *P*+, *P*+ is 25 cents lower than *P*++ and *P*++ is 25 cents lower than (*P*+1)--. Call the elements in the set N^{+}_{MIDI} the extended MIDI pitch.

In definition 5, the four extended MIDI pitch *P*--, *P*-, *P*+ *P*++ identified in the subjective hearing and located around the MIDI pitch *P* shown in Figure I are tried to be used to replace the traditional MIDI pitch *P*, with the purpose of improving the expression accuracy of the pitch. The final change result after replaced with the harmonics under the estimation models 1~5 with the MIDI pitch as the estimated result can be used to confer the extended MIDI pitch of the No. *t* frame of the singing signal under the estimation model 1 (namely, the sampling rate=44,000 Hz).

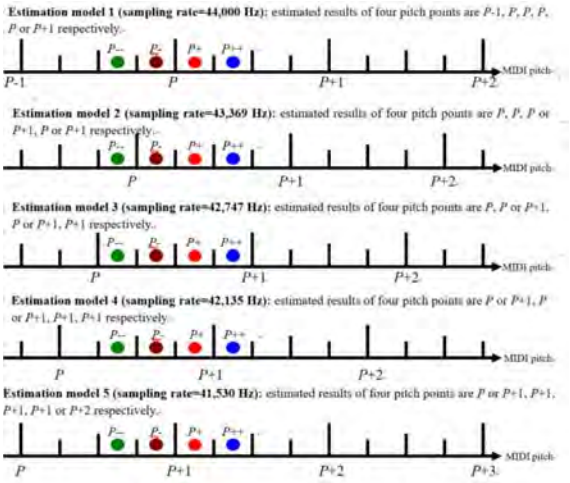

FIGURE I. ESTIMATION OF PITCH OF FOUR DIFFERENT PITCH POINTS LOCATED AROUND THE MIDI PITCH P UNDER FIVE PITCH MODELS

Table III presents the relationship between the result of four pitch points after executing the five estimation models and harmonics replacement and the extended MIDI pitch as shown in Figure I. Lines 2~8 in the Table describe the extended pitch of the MIDI pitch *P* in seven different situations when the estimated result of estimation model 1 is the MIDI pitch *P*. Each unit in columns 1~5 and lines 2~8 is called matching item. Among which, *Any* means arbitrary matching (namely, unconstrained matching); *Le*(*P*) means only matching the MIDI pitch no more than *P*; *Ge*(*P*+1) means only matching the MIDI pitch no smaller than *P*+1; *P* means only matching the MIDI pitch equaling *P*; *Ge*(*P*+2) means only matching the MIDI pitch no smaller than *P*+2.

TABLE III. THE RELATIONSHIP AMONG THE ESTIMATIMATION PITCH VALUE (INTEGER) BASED ON THE FIVE MODELS AND EXTENDED MIDI PITCH

Model 1 (sampling rate= 44,000)	**Model 2 (sampling rate= 43,369)**	**Model 3 (sampling rate= 42,747)**	**Model 4 (sampling rate= 42,135)**	**Model 5 (sampling rate= 41,530)**	**Extended MIDI pitch**
P	*Any*	*Any*	*Le*(*P*-1)	*Any*	*P*--
P	*Any*	*Le*(*P*)	*P*	*Any*	*P*--
P	*Any*	*Any*	*Any*	*Le*(*P*)	*P*--
P	*Any*	*Le*(*P*)	*Ge*(*P*+1)	*Any*	*P*-
P	*Le*(*P*)	*Ge*(*P*+1)	*Any*	*Any*	*P*+
P	*Any*	*Any*	*Any*	*Ge*(*P*+2)	*P*++
P	*Ge*(*P*+1)	*Any*	*Any*	*Any*	*P*++

For example, contents in line 5 in Table III are *P*, *Any*, *Le*(*P*), *Ge*(*P*+1), *Any*, *P*-. It points out that the name of the extended MIDI pitch under the vocal singing estimation model 1 is *P*-, if:

1) the estimated value of the No. *t* frame under the estimation model 1 is the MIDI pitch *P*; and

2) the estimated value of the No. *t* frame under the estimation model 2 is any MIDI pitch; and

3) the estimated value of the No. *t* frame under the estimation model 3 is matched no larger than the MIDI pitch*P*; and

4) the estimated value of the No. t frame under the estimation model 4 is matched no smaller than the MIDI pitch P+1; and

5) the estimated value of the No. t frame under the estimation model 5 is any MIDI pitch P.

Based on Table III it is not difficult for us to give the estimation algorithm for the extended MIDI pitch of the signal frame as stated in Algorithm 2.

Algorithm 2: Extended MIDI pitch estimation of the signal frame

Input: the output of Algorithm 1, i.e. the 5-tuple sequence: $<p_{r_1}(0), p_{r_2}(0), p_{r_3}(0), p_{r_4}(0), p_{r_5}(0)>$, $<p_{r_1}(1), p_{r_2}(1), p_{r_3}(1), p_{r_4}(1), p_{r_5}(1)>$, …, $<p_{r_1}(N), p_{r_2}(N), p_{r_3}(N), p_{r_4}(N), p_{r_5}(N)>$

Output: extended MIDI pitch sequence of the signal frame of the singing signal HS_1: $AP_0, AP_1, \ldots, AP_N$

Algorithm description:

For $t = 0, \ldots, N$ *do*{

AP_t="?"; /* the initial value of AP_t is put with the no-pitch candidate marker "?"*/

If $(p_{r_1}(t) = p_{r_5}(t))$ *or* $(p_{r_1}(t) - p_{r_4}(t) \geq 1)$ *or* $(p_{r_1}(t) \geq p_{r_3}(t)$ *and* $p_{r_1}(t) = p_{r_4}(t))$

then AP_t="P--"; /* assume $p_{r_1}(t)$ is valued P */

If $p_{r_3}(t) \leq p_{r_1}(t)$ *and* $p_{r_4}(t) - p_{r_1}(t) \geq 1$

then If AP_t="?" *then* AP_t="P-" *else* AP_t="P--|P-";

If $p_{r_2}(t) \leq p_{r_1}(t)$ *and* $p_{r_3}(t) - p_{r_1}(t) \geq 1$ *then*{

If AP_t="?" *then* AP_t="P+" ;

If AP_t="P--" *or* AP_t="P--|P-" *then* AP_t="P-";

If AP_t="P-" *then* AP_t="P-|P+"}

If $(p_{r_2}(t) - p_{r_1}(t) \geq 1)$ *or* $(p_{r_4}(t) - p_{r_1}(t) \geq 2)$ *then*{

If AP_t="?" *then* AP_t="P++" ;

If AP_t="P--" *or* AP_t="P--|P-" *then* AP_t="P--|P++";

If AP_t="P-" *or* AP_t="P-|P+" *then* AP_t="P+";

If AP_t="P+" *then* AP_t="P+|P++"}

Output AP_t}

There are two possible situations for each extended MIDI pitch in the extended MIDI pitch sequences output in algorithm 2. Generally, we uniformly denote the extended MIDI pitch of the signal frame t as P_t ($=x_t y_t | x_t y'_t$), among which, $x_t \in N_{\text{MIDI}}$ and $y_t, y'_t \in \{++, +, -, --\}$. N_{MIDI} is the integer set of the MIDI pitch. When $y_t = y'_t$, it means P_t is the only extended MIDI pitch; otherwise, when $y_t \neq y'_t$, it means P_t is the twin-candidate extended MIDI pitch. $x_t y_t | x_t y'_t$ means $x_t y_t$ or $x_t y'_t$.

Figure II shows the basic workflow of estimation of the extended MIDI pitch of a singing signal frame. Among which, the estimation method of the MIDI pitch of the signal frame with variable sampling rate is as stated in Definition 3. The harmonics replacement and estimation of the extended MIDI pitch of the signal frame are respectively described in Algorithms 1 and 2. Pitch correction refers to adjusting the possible wrong results caused by the error in the rounding operation in the equation (2).

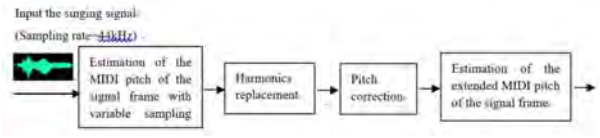

FIGURE II. WORKFLOW OF ESTIMATION OF EXTENDED MIDI PITCH OF THE SINGING SIGNAL FRAME

III. EXPERIMENT AND EVALUATION

A. Music Collection and Evaluation Criteria

We use test data (including 38 human singings with duration of 1154s and their ground truth files) provided by Emilio Molina [19] in http://www.atic.uma.es/ismir2014 singing/for comparison. The 38 ground truth files labeled by musicians artificially in the data can be deemed as 38 evaluation criteria of transcribed actual melodies estimated in our methods.

B. Result Comparison and Discussion

Table IV is subject to the criteria of ground truth to estimate the result of transcribed melody that is estimated by using our method. Figure III shows the result that is based on the test result of Emilio Molina [19] mixed with our approach for comparison. Where, the raw pitch accuracy is the percentage of voiced frames where the pitch estimation is correct, voicing recall is the percentage of voiced frames in the reference that classified as voiced by the algorithm, voicing false alarm is the percentage of unvoiced frames in the reference that are classified as voiced by the algorithm.

TABLE IV. THE ASSESSMENT RESULTS OF 38 SINGING FRAGMENTS

Total time of singing	raw pitch accuracy	voicing recall	voicing false alarm
1154 secs	86.34	85.87	22.11

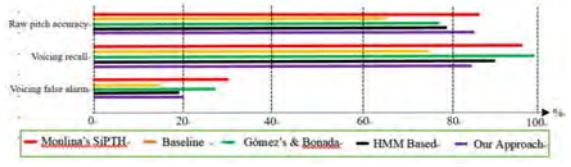

FIGURE III. DETAILED PERFORMANCE EVALUATION ON TRANSCRIBED MELODY OD OUR SINGING TRANSCRIPTION SYSTEM, MONLINA'S SiPTH IN [19], TIITANIEMI'S BASELINE APPROACH IN [18], GóMEZ'S TRANSCRIPTION SCHEME IN [20] AND RYYNäNEN'S APPROACH in [17]

In terms of the estimation of fundamental frequency of signal frames, the raw pitch accuracy achieved by our algorithm is up to 86.35%, very close to the best result achieved by SiPTH system [19] of Monlina et al. The approach of combination of the YIN algorithm with better fundamental frequency estimation

accuracy and the Hysteresis Defined on the Pitch-Time Curve put forward by Monlina et al is used in the SiPTH system, but our algorithm can only control the pitch error within 25-50 cents and the effective range confined within three octaves [F2, E5]. The computation complexity of our fundamental frequency estimation approach is $o(n\log_2 n)$ and that of the YIN algorithm is $o(n^2)$.

IV. CONCLUSION

The approach of Monlina's SiPTH system [19] is difficult to use in many existing practical query by humming system although it achieves the best raw pitch accuracy in the experiment. It is because the YIN algorithm used in the system requires the computation complexity of $n(n^2)$. By comparison, our approach becomes the best option not only because of its good raw pitch accuracy but also the less computation complexity of $n(\log_2 n)$.

ACKNOWLEDGMENT

The authors are grateful to E. Molina for providing the results of the scheme developed in [19] for comparison. This work is supported by the National Natural Science Foundation of China under Grant No. 61272238.

REFERENCES

[1] M. Lesaffre, M. Leman, B. De Baets, and J. Martens, "Methodological considerations concerning manual annotation of musical audio in function of algorithm development", in *Proc. 5th Int. Conf. Music Inf. Retrieval ISMIR*, 2004, pp.64-71.

[2] De Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music." *J. Acoust. Soc. Amer.*, vol. 111.no.4. pp. 1917-1930, 2002.

[3] H. Kobayashi and T. S. Himamura, "A weighted autocorreleation method for pitch extraction of noisy speech", in *Proc. of International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'00)*, 2000, Vol. 3, pp.1307-1310.

[4] L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection", *IEEE Transactions on Acoustics, Speech, And Signal Processing*, Vol.25, Issue.1, pp.24-33, Feb. 1997.

[5] E. Dorken and N. S. Hamid, "Improved musical pitch tracking using principal decomposition analysis", in *Proc. of International Conference on Acoustics, Speech, and Signal Processing, (ICASSP'94)*, 1994, pp. II/217-II/220.

[6] W. J. Pielemeier, G. H. Wakefield, "Time-frequency and time-scale analysis for musical transcrioption" in *IEEE Symp. on Signal Processing.(IEEE-SP'92))*, 1992, pp.421-424.

[7] Judith C. Brown, "Calculation of a constant Q spectral transform", *J. Acoust. Soc. Amer*. vol. 89, Issue 1, pp.425–434, Jan. 1991.

[8] Judith C. Brown et al. "A high resolution fundamental frequency determination based on phase changes of the Fourier transform", *J. Acoust. Soc. Amer*. vol. 94, Issue 2, pp. 662-667, August 1993.

[9] Judith C. Brown, "Frequency ratios of spectral components of musical sounds", J. Acoust. Soc. Amer., vol. 99, Issue 2, pp. 1210-1218, Sept. 1996.

[10] Adriano Mitre, Marcelo Queiroz, Regis R.A.Faria, "Accurate and Efficient Fundamental Frequency Determination from Precise Partial Estimates", in *Proc. of the 4th AES Brazil Conference*, May 2006, pp. 113–118.

[11] Wei-Ho Tsai and Hsin-Chieh Lee, "Automatic Evaluation of Karaoke Singing Based on Pitch, Volume, and Rhythm Features", *IEEE Transactions on Audio, Speech, And Language Processing*, Vol.20, No.4, May 2012, 1233-1243.

[12] Yin Feng and Wentao Wang, "Research on humming to MIDI by man-machine interaction", *J. of Computational Information System*, vol. 9, no.10, May 2013, pp.3827-3835.

[13] Stephen Zahorian and Hongbing Hu, "A spectral/temporal method for robust fundamental frequency tracking", *J. Acoust. Soc. Amer*., vol. 123. no.6, 2008, pp. 4559-4571.

[14] Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by humming: musical information retrieval in an audio database", in *Proc. of ACM International Conference on Multimedia*, 1995, pp. 231-236.

[15] T. R. Black and K. D. Donohue, "Pitch determination of music signals using the generalized spectrum", in *Proc. of the IEEE Southeast Conference, Nashville, USA*, 2000, pp.104-109.

[16] [16]W. Keige, T. Herbst, and T. Niesler, "Ecplicit transition modelling for automatic singing transcription," *J. New Music Res.*, vol. 37, no. 4, pp. 311-324, 2008.

[17] M. Ryynänen, A. P. Klapuri, "Modelling of Note Events for Singing Transcription." in P*roc. ISCA Tutorial and Res. Workshop on Statist. Percept. Audio Process. SAPA, Jeju, Korea*, Oct. 2004.

[18] T. Viitaniemi, A. Klapuri, and A. Eronen, "A probabilistic model for the transcription of single-voice melodies," *in Proc. of Finnish Signal Process. Symp. (FINSIG'03)*, 2003, pp.5963-5957.

[19] Emilio Molina et al, "SiPTH: Singing Transcription Based on Hystreresis Defined on the Pitch-Time Curve", I*EEE/ACM Transactions on Audio, Speech, And Language Processing*, Vol.23, No.2, Feb. 2015, 252-261.

[20] Emilia Gómez and Jordi Bonada, "Towards Computer-Assisted Flamenco Transcription: An Experimental Comparison of Automatic Transcription Algorithms as Applied to A Cappella Singing", *Computer Music Journal*, Vol.37, Issue 2, 2013, 73-90.

[21] R. J. McNab, L. A. Smith, and I. H. Witten, "Signal processing for melody transcripton", in *Proc. 19th Australasian Comput. Sci. Conf*., 1996, vol.18, no.4, pp 301-307.

[22] G.Haus and E.Pollastri, "Audio front end for query-by-humming systems", in *Proc. 2nd Inf. Retrieval* (*ISMIR*), 2001, pp.65-72.

[23] L P. Clarisse, J.P. Martens, M.Lesaffre, B.D. Baets, H.D. Meyer, and M. Leman, "An auditory model based transcriber of singing sequences," in *Proc. 3rd Int. Conf Music Inf. Retrieval ISMIR*, 2002, pp. 116-123.

[24] T. De Mulder, J. P. Martens, M. Lesaffre, M. Leman, B. D. Baets, H.D. Meyer, "An auditory model based transcriber of vocal queries," in *Proc. 4th Int. Conf. Music Inf. Retrieval ISMIR*, 2003.

[25] Cheng-Yuan Lin and Jyh-Shing Roger Jang, "Automatic Phonetic Segmentation by Score Predictive Model for the Corpora of Mandarin Singing Voices", *IEEE Transactions on Audio, Speech, And Language Processing*, Vol.15, No.7 Sept. 2007, 2151-2159.

[26] Chee-Chuan Toh, Bingjun Zhang, Ye Wang, "Multiple-Feature Fusion Based Onset Decetion for Solo Singing Voice" in *Proc. of ISMIR, Philadeiphia, PA. USA*, 2008, pp.515-520.

[27] G. Hinton, L. Deng, D.Yu, G.Dahl, A. Mohamed, N. Jaitly, A. Senior, V.Vanhouke, P.Nguyen, T.Sainath, and B.Kingsbury, "Deep neural networks for acoustic modeling in speech recognition", *IEEE Signal Processing Mag*, vol. 29, no6, pp.82-97, 2012.

[28] L. Krumhansl, *Cognitive Foundations of Musical Pitch, Oxford Psychology Series No*. 17, *New York Oxford, Oxford University Press*, 1990.

[29] Andrew Guillory et al. "User-Specific Learning for Recognizing a Singer's Intended Pitch", in *Proc. of the Twenty-Fourth AAAI Conference on Artificial Intelligence, Atlanta, Georgia, USA*, July 11-15, 2010, pp.960-966.

[30] S. Porter, "MUSIC: A Comprehensive Introduction", *Published by Excelsior Music Publishing Co. 15 West 44th St., New York, N.Y.* 10036, 1986.

[31] J. Salamon, J. Serra, and E. Gómez, "Tonal representations for music retrieval: From version identification to query-by-humming", *Int. J. Multimedia Inf. Retrieval*, pp.1-14, 2013.

[32] P. M. Brossier. "Automatic annotation of musical audio for interactive applications.," *Ph.D. dissertation, Centre for Digital Music, Queen Mary, Univ. of London, London, U.K.*, 2006.