Applied Acoustics 112 (2016) 70–78

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Main melody extraction from polyphonic music based on modified Euclidean algorithm

Weiwei Zhang [a,b], Zhe Chen [a], Fuliang Yin [a,*]

[a] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116023, China
[b] School of Information and Communication Engineering, Dalian Minzu University, Dalian 116605, China

ARTICLE INFO

Article history:
Received 22 December 2015
Received in revised form 26 April 2016
Accepted 27 April 2016
Available online 21 May 2016

Keywords:
Main melody extraction
Euclidean algorithm
Music information retrieval
Polyphonic music

ABSTRACT

Extracting main melody from polyphonic music is one of the most appealing and challenging tasks in music information retrieval (MIR). In this paper, a new melody extraction method based on a modified Euclidean algorithm (MEA) is proposed. Firstly, the instantaneous frequency is adopted to gain better frequency discrimination, and the frame-wise pitch candidates are estimated based on the modified Euclidean algorithm. Next, the candidate trajectories are formed using these potential candidates, and padded by the candidate one octave above or below if there is a gap at some isolated frames. Finally, the melodic contours are extracted using the melody smoothness and salience principle. The proposed modified Euclidean algorithm can deal with diverse coprime harmonic combinations, and work well at low computational cost and memory requirement. The experimental results show that the proposed method can extract main melody extraction effectively with few pitch candidates.

1. Introduction

Music is ubiquitous and can be easily accessed through Internet, smart phones, CD tracks and so on. As the core of western music, melody plays an important role in many applications, including query by humming (singing) [1,2], cover song identification [3], automatic music transcription [4], audio-to-score alignment [5], *etc*. However, extracting main melody from polyphonic music is technically challenging because two or more notes may sound simultaneously. Therefore, the extraction of main melody from polyphonic music signal has become an active research topic in music information retrieval community in the past decade.

Main melody extraction is also known as the predominant melody extraction or predominant fundamental frequency f_0 estimation of polyphonic music. The existing melody extraction approaches can mainly be divided into two categories [18]: the salience-based methods [6–9] and the source separation-based methods [10–12]. The source separation-based methods combine the source separation algorithms with the monophonic pitch tracking strategies. Durrieu et al. exploited non-negative matrix factorization to separate the melody from the mixture, and employed Viterbi algorithm to track the most probable fundamental frequency sequence [10]. Arora and Behera presented a predominant vocal melody estimation method by tracking various sources with the help of harmonic clusters, and then determined the predominant vocal source by its harmonic strengths [12]. Most of these melody extraction methods detect pitches in some plausible pitch ranges. However, since melody pitches vary in a wide range, it is difficult to give an appropriate pitch range if no prior knowledge is given. To address this problem, Hsu et al. proposed a trend estimation algorithm for singing pitch estimation, which can determine the pitch ranges adaptively and improve singing pitch detection [11]. Most of the source separation-based methods require large memory to store high resolution STFT spectrogram and harmonics of each source. Moreover, their performances are constrained by whether the sources are explicitly separated from the polyphonic excerpts.

The salience-based methods construct various salience functions, and select the most dominant contours as the final main melody. In 2004, Goto studied melody line and bass line extractions from real-world polyphonic music, and proposed a predominant f_0 estimation method by a maximum posteriori probability (MAP) algorithm [6]. From then on, some other salience-based methods have also been proposed, where different salience functions are exploited to measure the dominance of main melody. Some researchers constructed salience functions based on the harmonic amplitudes [6,7,13,14]. Motivated by the fact that the fundamental and harmonic frequencies of a given note exhibit

* Corresponding author.
E-mail addresses: zhangww@dlnu.edu.cn (W. Zhang), zhechen@dlut.edu.cn (Z. Chen), flyin@dlut.edu.cn (F. Yin).

http://dx.doi.org/10.1016/j.apacoust.2016.04.023
0003-682X/

specific deviations from the equal tempered scale, Degani et al. exploited harmonic frequency differences for the salience function [15]. Sub-harmonic summation is widely used to measure the salience of the frame-wise pitches [16]. However, the fundamental and secondary harmonics are often distorted by the bass, percussion or other simultaneously sounded pitched accompaniments. In 2011, Dressler proposed a pitch estimation method based on the pair-wise evaluation of spectral peaks [9]. Dressler's method can solve the missing fundamental frequency problem. However, its assumption that the spectral peak pairs detected are successive (odd) harmonics is not always the case in practice. In addition, according to Dressler's pitch estimation method, two or four candidates are generated for each spectral peak pair. Thus, the proposed main melody extraction method intrinsically requires a sophisticated audio streaming scheme [17].

To remedy the missing fundamental frequency problem, a new melody extraction method based on a modified Euclidean algorithm is proposed in this paper. The instantaneous frequency spectrogram derived from the phase information is adopted to estimate the sinusoidal components of polyphonic music. Then, the Euclidean algorithm for calculating the greatest common divisor of two natural numbers is generalized to the case with float numbers, named as the modified Euclidean algorithm (MEA) herein, and is exploited to estimate the pitch candidates in each frame. Next, the frame-wise pitch candidates are grouped into candidate melody contours, and the isolated and spurious ones are eliminated. Finally, the melodic contours are identified according to the temporal smoothness. The proposed method can deal with the missing fundamental frequency problem. Moreover, it has less memory requirement, and does not rely on any prior information, which make it suitable for various genres of polyphonic music.

The remainder of this paper is organized as follows. Section 2 gives the polyphonic music model and briefly introduces the instantaneous frequency spectrum. Section 3 elaborates the proposed method, including the modified Euclidean algorithm, the pitch candidate estimation, creation of the pitch contours, and the selection of the most possible contours to form the main melody trajectory. Section 4 presents some experimental results and discussions. Finally, some conclusions are drawn in Section 5.

2. Background knowledge

2.1. Problem formulation

There are two basic assumptions concerning the main melody extraction task: (1) the salience principle, i.e., the main melody line consists of the notes with the highest intensity; (2) the smoothness principle, i.e., melodies are often smooth in terms of note-frequency intervals [7,8]. Polyphonic music in time domain can be simply modeled as [18]

$$y(t) = x(t) + n(t) \tag{1}$$

where $y(t), x(t)$ and $n(t)$ are the observed polyphonic music signal, the target monophonic melody signal, and the additive accompaniment, respectively.

The task of main melody extraction is to estimate the pitch sequence $\widehat{f}$ from the audio recording of a polyphonic music [18], i.e.,

$$\widehat{f} = \arg\max_{f} \sum_{\tau} s_y(f_\tau, \tau) + c(f) \tag{2}$$

where $\sum_\tau s_y(f_\tau, \tau)$ is the pitch salience function calculated over the observed $y(t)$ and is tackled as the multi-pitch representation, and $c(f)$ represents the temporal constraints and is addressed by employing tracking strategies.

2.2. Instantaneous frequency estimation

Melody is most relevant to the sinusoidal components of the audio recordings. Yet, it is a challenge to achieve high enough frequency resolution while tracking the rapid changing of music. Only pitches with less than half semitone deviation from the ground truth are considered to be correct. Hence, instantaneous frequency (IF) is often adopted to obtain more accurate sinusoidal frequency estimation. There are many methods for the IF and amplitude estimation based on Fourier analysis. Keiler and Marchand found that the methods based on the phase information give the best results [19]. Goto employed the Abe IF method [20], while Dressler and Salamon used the average of phase vocoder and the Charpentier IF method [21] to obtain more stable IF measures. The three methods are compared using the ground truth of fundamental and corresponding harmonic components on ADC2004 melody extraction dataset. The evaluation results show that the Charpentier and Abe methods obtain similar results, which outperform phase vocoder in terms of minimum mean square error criterion. Therefore, Abe's method is selected to estimate the IF in this paper. Suppose that the short-time Fourier transform of signal $x(t)$ is denoted as $X(\omega,t) = a(\omega,t) + jb(\omega,t)$, where $a(\omega,t)$ and $b(\omega,t)$ are the real and imaginary parts of $X(\omega,t)$, respectively. The IF at (ω,t) is defined as [20]

$$\lambda(\omega,t) = \frac{\partial}{\partial t}\arg[X(\omega,t)] \tag{3}$$

where arg[·] denotes the argument function of a complex function. Then $\lambda(\omega,t)$ is calculated by

$$\lambda(\omega,t) = \frac{a(\omega,t)\frac{\partial b(\omega,t)}{\partial t} - b(\omega,t)\frac{\partial a(\omega,t)}{\partial t}}{a^2(\omega,t) + b^2(\omega,t)} \tag{4}$$

The amplitude of the IF at (λ_0,t) is calculated by

$$g(\lambda_0,t) = \lim_{\Delta\lambda\to 0}\int_{\lambda_0 \leqslant \lambda(\omega,t) \leqslant \lambda_0+\Delta\lambda} |X(\omega,t)|d\omega \tag{5}$$

3. Modified Euclidean algorithm based main melody extraction

In general, each note is composed of the fundamental frequency and harmonic components. However, the fundamental frequency components are absent occasionally in some arias (such as opera). Thus, it is not an ideal approach to locate the fundamental frequency directly. Fortunately, there are abundant harmonics in music signal, which are the integer multiples of the fundamental frequency. Some researchers have demonstrated that human auditory system can still perceive the pitch even though the fundamental component is absent [22]. Moreover, the auditory system prefers to perceive the pitch from adjacent harmonics called *residue pitch* [23]. How to find the *residue pitch* when the fundamental component is absent is a key issue for the main melody extraction. To address this problem, a modified Euclidean algorithm is proposed herein.

The proposed method consists of three blocks: sinusoidal estimation, multi-pitch representation, and melodic trajectory tracking, as depicted in Fig. 1 (following the salience-based melody extraction framework [18]). The sinusoidal estimation is used to find the frame-wise sinusoidal components, multi-pitch representation is employed to obtain the frame-wise pitch candidates, and the melodic trajectory tracking is to construct pitch trajectories and select the best one as the main melody output. Following the similar steps for sinusoidal extraction as others, the sinusoidal estimation is implemented sequentially by short-time Fourier transform, spectral peak search and IF correction (by Abe method). In

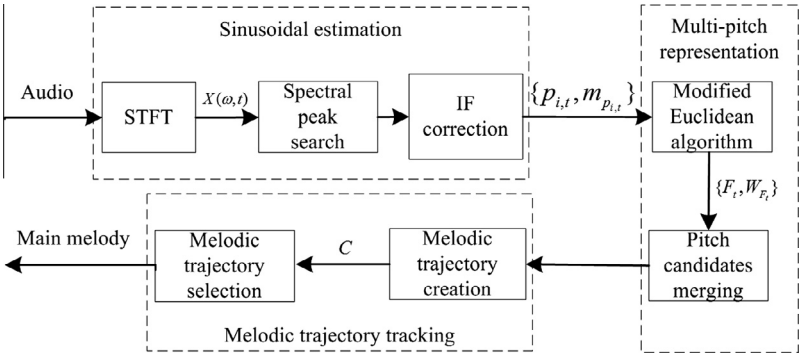


Fig. 1. Block diagram of the proposed melody extraction algorithm.

the following subsections, the multi-pitch representation and melodic trajectory tracking will be described in details.

3.1. Multi-pitch representation

3.1.1. Modified Euclidean algorithm

Euclidean algorithm is widely used to calculate the greatest common divisor (GCD) of two natural numbers [24]. It is based on the principle that the greatest common divisor of two numbers does not change if the larger number is replaced by its difference with the smaller number. Moreover, the GCD can be obtained by iteratively replacing the greater by the remainder of the greater divided by the smaller until the remainder is equal to zero.

Suppose that a and b, $(a > b)$ are two natural numbers. The remainder r_k is calculated as

$$r_k = a - qb \tag{6}$$

where $r_k, a, q, b \in \mathbb{N},\ 0 \leqslant r_k < b$.

If r_k is greater than zero, then $a = \max\{b, r_k\},\ b = \min\{b, r_k\}$, and compute r_k again. Repeat the procedure until r_k is equal to zero. Thus, b in the last iteration is the greatest common divisor.

The fundamental frequency can be regarded as the *greatest common divisor*, also called analogy greatest common divisor herein, of the harmonics. To obtain the fundamental frequency of a note, we generalize the Euclidean algorithm to float numbers. Assume that x and y are two float numbers and $0 < x < y$. The *analogy greatest common divisor* of x and y can be calculated by evaluating $r(x,y)$, defined as

$$r(x,y) = \left| \frac{y}{x} - \left[\frac{y}{x}\right] \right| \tag{7}$$

where $[\cdot]$ denotes the round operator which outputs the nearest integer of the input argument, and $|\cdot|$ represents the absolute operator.

$r(x,y)$ is the Euclidean distance between $\frac{y}{x}$ and $\left[\frac{y}{x}\right]$. If x is the analogy greatest common divisor, $\frac{y}{x}$ would be very close to $\left[\frac{y}{x}\right]$, and thus $r(x,y)$ approaches zero, indicating that x and y are in harmonic relationship. If they are not in harmonic relationship, $r(x,y)$ would be larger, but still smaller than or equal to 0.5. So $r(x,y)$ is a measure indicating whether the two frequencies are in harmonic relationship.

If $r(x,y)$ is smaller than the threshold ς, the analogy greatest common divisor is obtained by

$$\gcd(x,y) = \frac{x+y}{1+h_y} \tag{8}$$

where $h_y = \left[\frac{y}{x}\right]$.

The analogy greatest common divisor is defined according to Eq. (8) rather than x to obtain the mean value and reduce estimation error.

3.1.2. Modified Euclidean algorithm based pitch candidates calculation

Each note is composed of its fundamental and harmonic components. The fundamental frequency is perceived as pitch. All of the harmonic components are integral multiples of the fundamental frequency. However, the fundamental frequency is sometimes concealed by some special singing strategies or masked by other concurrent sounds. In these situations, human can still perceive the virtual pitch, i.e., the *greatest common divisor* of the harmonics [22]. Motivated by this fact, the perceived pitches are computed by the proposed modified Euclidean algorithm (MEA), which utilizes the sinusoidal frequencies.

As shown in Fig. 2, each spectral peak is described by two parameters, i.e., the frequency $p_{l,t}$ and amplitude $m_{l,t},\ l = 1, \ldots, n_p$, where n_p is the number of peaks. $p_{i,t}$ and $p_{j,t} (p_{i,t} < p_{j,t})$ are assigned to the x and y in Eq. (7), respectively. Then, the pitch candidate $f_{k,t}$ is obtained by using the modified Euclidean algorithm.

There are some spurious candidates (from the non-melodic sources), harmonic and sub-harmonic ones besides the fundamental frequency. Thus, the candidates are ranked in the descending order according to their weights. Motivated by the salience principle, the pitch candidates derived from the peaks both with greater amplitudes are more likely to be the virtual pitch. The weight of $f_{k,t}$ is defined as

$$w_{f_{k,t}} = m_{i,t} m_{j,t} \tag{9}$$

where $m_{i,t}$ and $m_{j,t}$ are the amplitudes of the ith and jth peaks, respectively.

The pitch calculation algorithm is summarized in detail in Algorithm 1.

Note: Eq. (7) is just one of the iteration steps in Algorithm 1, so for the first iteration, x and y are equal to the smaller frequency $p_{i,t}$

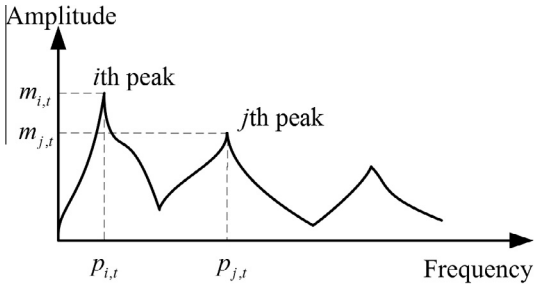


Fig. 2. Spectrum at frame t.

and greater frequency $p_{j,t}$, respectively. From the second iteration, x is the same, but y is equal to the modulus after division $\frac{y}{x}$, as depicted in step (3) of Algorithm 1.

3.1.3. Pitch candidates merging

The L spectral peaks will generate $\frac{1}{2}L(L-1)$ peak pairs in a frame, then $\frac{1}{2}L(L-1)$ pitch candidates are obtained by the MEA. However, some of the pitch candidates may fall out of the pitch range, or may be very close, e.g. the frequency difference is less than half semitone representing the same pitch. The candidates in the former case should be discarded, while the ones belonging to the latter should be merged together for later processing.

The task of pitch candidates merging is to combine these closely located pitches as one, and calculate the mean of the frequencies as the final pitch. The largest weight calculated according to Eq. (9) is selected as the final weight. Pitch candidate merging is implemented as follows. For $\forall f_{k,t} \in F_t$, the emerged frequency is computed as

$$\bar{f}_{k,t} = \frac{1}{\Omega}\sum_{f\in\Omega} f \tag{10}$$

where $\Omega = \{f \,|\, |f - f_{k,t}| < 50 \text{ cents}\}$.

After pitch candidates merging, the number of possible candidates is reduced greatly, but still more than one.

3.2. Melodic trajectory tracking

At the multi-pitch representation stage, pitch candidates are obtained on a per frame basis. Some candidates may be very close to the ones in adjacent frames, while the others may not. Melodic trajectory tracking aims to group these f_0 candidates into melody fragments, and eliminate the spurious ones. It includes melodic trajectory creation and selection, which will be described in detail as follows.

3.2.1. Melodic trajectory creation

During melodic trajectory creation stage, the following two tasks must be fulfilled [7]: (1) the short time gaps in the pitch trajectory do not split a single contour into several contours; (2) the candidates which are isolated or have few similar values in contiguous frames are considered to be spurious ones. To do so, the candidates are grouped into the melodic contours by Marolt method [25].

As aforementioned, some pitch candidates are derived from the peak pairs. There are combinations of non-coprime pairs, e.g. the 2nd and 4th harmonics, thus it is possible to obtain the multiples of f_0. Many harmonics detected are not in perfect multiple relationships, so there are still some sub-multiple candidates. Both multiples and sub-multiples of the actual pitch are so-called octave errors [18]. However, they can be used to complement the melodic contours. Assume that $f_{c_j,t}$ and $f_{c_k,t}$ are the pitches of contour c_j and c_k at time t, respectively. The complement of melodic contour c_j is accomplished as

$$f_{c_j,t} = \begin{cases} \frac{1}{2}f_{c_k,t} & s_1 = \emptyset,\ s_2 \neq \emptyset \\ 2f_{c_k,t} & s_1 = \emptyset,\ s_3 \neq \emptyset \\ f_{c_j,t} & s_1 \neq \emptyset \end{cases} \tag{11}$$

where $\emptyset$ denotes the null set, and s_1, s_2, s_3 are the fundamental frequency, second harmonic and subharmonic components, respectively, and defined as

$$s_1 = \left\{ f_{c_j,t} \,\middle|\, 1200\left|\log_2 \frac{f_{c_j,t}}{f_{c_j,t-1}}\right| < 100 \text{ cents} \right\} \tag{12}$$

$$s_2 = \left\{ f_{c_k,t} \,\middle|\, 1200\log_2\left|\frac{f_{c_k,t}}{2f_{c_j,t-1}}\right| < 100 \text{ cents} \right\} \tag{13}$$

$$s_3 = \left\{ f_{c_k,t} \,\middle|\, 1200\left|\log_2 \frac{2f_{c_k,t}}{f_{c_j,t-1}}\right| < 100 \text{ cents} \right\} \tag{14}$$

$s_i = \emptyset,\ i \in \{1,2,3\}$ implies that there is no candidate belonging to the corresponding contour. If $s_1 = \emptyset$, there is no estimated pitch candidate belonging to melodic contour c_j at time t, but the contour can be complemented by the candidate one octave below or above.

3.2.2. Melodic trajectory selection

In the melodic trajectory selection stage, the impossible contours are gradually removed from the melodic contours to further reduce the false selection risk. The aim of selection algorithm is to select the appropriate melodic trajectories among several overlapped candidate trajectories, sometimes only one.

Algorithm 1. Computation for the parameters of pitch candidates: $\theta_t = \{F_t, W_{F_t}\}$

For $t = 1, \ldots, p$, n_p is the number of peaks at time t,
For $i,j = 1, \ldots, n_p$ $i \neq j$, do
$x = p_{i,t}, y = p_{j,t}$, where $p_{i,t}$ and $p_{j,t}$ are the ith and jth spectrum peak frequencies at frame t, respectively.
(1) $y \leftarrow \max(x,y), x \leftarrow \min(x,y)$;
(2) compute $r(x,y) = \left|\frac{y}{x} - \left[\frac{y}{x}\right]\right|$;
(3) if $r(x,y) \geqslant \varsigma$
then $z = \text{mod}(y,x), y = z$, where $\text{mod}(y,x)$ returns the modulus after division $\frac{y}{x}$, then go to step (1);
if $r(x,y) < \varsigma$
then the pitch candidate is given according to Eq. (8), and $f_{k,t}$ is assigned with a corresponding weight
$w_{f_{k,t}} = m_{p_{i,t}} m_{p_{j,t}}$. If there are still peak pairs which haven't been evaluated, select another pair and go back
to step (1);
(4) Output the pitch candidates and their related weights, i.e., $\theta = \{F_t, W_{F_t}\}$, where $F_t = \{f_{k,t}\}$ and $W_{F_t} = \{w_{f_{k,t}}\}$.
End for
End for

The contours shorter than 120 ms are eliminated based on the fact that western music tends to have notes rarely shorter than 150 ms [26]. The margin is to reduce the risk of deleting the melody contours. Then the contours c_i and c_j are linked together to construct long contours if the following three conditions are satisfied:

(1) $|t_{j,start} - t_{i,end}| < 50$ ms;
(2) $\left|f_{c_j,t_{j,start}} - f_{c_i,t_{i,end}}\right| < 7$semitones;
(3) $\left|f_{c_j,t_{j,start}} - f_{c_i,t_{i,start}}\right| \leqslant \left|f_{c_k,t_{k,start}} - f_{c_i,t_{i,start}}\right|,\ k = 1, \ldots, C$

where $t_{j,start}$ represents the start time of contour c_j, $t_{i,end}$ the end time of contour c_i $f_{c_j,t_{j,start}}$ the frequency of the contour c_j at the start time, $f_{c_i,t_{i,end}}$ the frequency of the contour c_i at the end time, and C the number of contours.

The long contours whose lengths are shorter than 200 ms are removed due to the fact that they seem to be isolated contours. Then the long contours are broken into the original short ones. For the regions with no overlapped contours, the only contour is

selected as the final one. For the overlapped contours, the one nearest to the adjacent contours is chosen as the final one in accordance with the smoothness principle.

3.3. Computational complexity and memory requirement

Computation of the proposed melody extraction method is mainly concentrated on two stages: STFT and MEA. STFT can be implemented through FFT whose computational complexity is on the order of $O(\frac{1}{2}N\log_2 N)$. The FFT frequency discrimination can be improved through the instantaneous frequency method, yielding smaller N. Given top L peaks for each frame, there are $\frac{1}{2}L(L-1)$ candidate pitches calculated by the MEA. However, according to the observations on polyphonic music, a noticeably large number of these candidates are very close because they originate from the harmonics of the same source. Thus, the candidate pitches whose iteration time is larger than three is set to zero to reduce iteration computational cost.

STFT spectrogram requires heavy memory. Fortunately, the proposed method reduces memory requirement in two ways: (1) IF estimation helps reduce quantization error with a shorter FFT length; (2) the memory used to store STFT spectrogram can be released after multi-pitch representation, as only the frame-wise pitch candidates (only one pitch for each frame is selected to illustrate the melody extraction performance in this paper) are delivered for melody contour creation and selection.

4. Evaluations and discussions

To verify the validity of the proposed method, some evaluation experiments are carried out. In the following subsections, the evaluation metrics, evaluation datasets, and experimental results are given explicitly.

4.1. Evaluation metrics

The proposed melody extraction method is evaluated in accordance with the MIREX melody extraction task. Two metrics, raw pitch accuracy (RPA) and raw chroma accuracy (RCA), are used to evaluate the performance of the proposed melody extraction method. These two metrics are defined as [27]

$$RPA = \frac{\#\{voiced\ true\ positive\ pitches\}}{\#\{voiced\ frames\}} \tag{15}$$

$$RCA = \frac{\#\{voiced\ true\ positive\ chromas\}}{\#\{voiced\ frames\}} \tag{16}$$

4.2. Evaluation datasets

The proposed method is evaluated on the ADC2004, the MIREX05 train and MIR-1K datasets. ADC2004 is provided by the Music Technology Group of the Pompeu Fabra University. There are 20 polyphonic musical recordings in the dataset, including MIDI, Jazz, Pop and Opera, where each clip lasts around 20 s. The STFT frame size and hop size of the analysis windows are 2048 and 256, respectively. As the sampling rate is 44.1 kHz, each frame is about 46 ms (=2048/44,100) long and the hop step is about 5.8 ms (=256/44,100).

MIREX05 train dataset is collected by Graham Poliner and Dan Ellis (LabROSA, Columbia University), including 13 excerpts lasting from 24 to 39 s. The time interval for this dataset is 10 ms, so the hop size is 441 (=44,100 × 0.01). The other parameters are the same as the setups of ADC2004.

MIR-1K contains 1000 song clips extracted from 110 karaoke songs recorded at 16 kHz sampling rate with 16-bit resolution. The duration of each clip ranges from 4 to 13 s, and the total length of the dataset is 133 min. These songs were selected from 5000 Chinese pop songs and sung by 19 amateur singers. The hop size is 160(=16,000 × 0.01). The other parameters are the same as those of ADC2004.

4.3. Performance of the proposed modified Euclidean algorithm

4.3.1. Validity of the proposed MEA

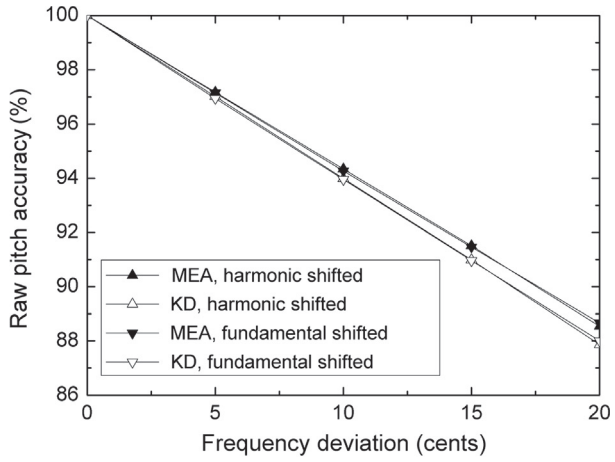
The pitch estimation methods can be divided into two categories: (1) only peak frequency based method; (2) both peak frequency and amplitude based method. Dressler's pair-wise evaluation (denoted as KD for conciseness) is one of the best frequency based methods [9]. The proposed method also relies exclusively on the peak frequencies, so these two pitch estimation performances are compared here. The pitch estimation accuracy is evaluated in two cases: the existing fundamental frequency and missing fundamental frequency cases.

There are two situations for the existing fundamental frequency case: (1) f_0 is accurate (e.g. 100 Hz), but the harmonics are not exactly its multiple times, which can be considered as the exact harmonic plus some shifting; (2) the harmonic frequencies are accurate (e.g. 200 or 300 Hz), while the fundamental frequencies are shifted. As the KD's method can only deal with the successive (odd) harmonics, so only the fundamental frequency (100 Hz) and second or third harmonic are used.

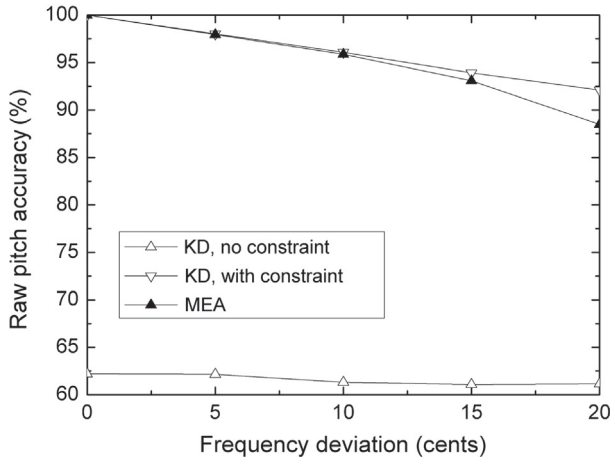
According to psychoacoustical experiments, the mistuned harmonic deviation is often set to be less than 20 cents. Thus, the shifted frequencies are randomly generated with standard deviation varied from 0 to 20 cents. The evaluation results when changing either fundamental frequency or harmonic are illustrated in Fig. 3(a), where $\varsigma = 0.15$. The hollow up triangle and down triangle indicate the raw pitch accuracy by KD's method, while the solid up triangle and down triangle represent the raw pitch accuracy by the MEA. KD's method can obtain four pitches for each pair at most, so in this simulation only the pitches within one semitone range of the ground truth are considered and randomly selected, which potentially reduces the error rate. It's obvious from Fig. 3(a) that the pitch estimation accuracies decrease with the increment of inharmonicity. Both of the two methods have similar sensitivity to the higher and lower peak frequencies. The proposed MEA gains slightly higher raw pitch accuracy than KD's method. In addition, KD's method can only deal with the successive (odd) situation, while the proposed MEA can deal with various other coprime combinations. Experimental results show that the raw pitch accuracy is robust among different combinations. There are two or four estimated potential pitches for each peak pair obtained by KD's method, which might make it very controversial in real world recordings, as no prior information about the melody is given. So a very sophisticated streaming scheme is needed to eliminate the spurious pitches.

For the missing fundamental frequency situation, three groups are taken into consideration: second–third, third–fifth, fourth–fifth harmonic pairs. In each group, either frequency of each combination varies with standard deviation ranging from 0 to 20 cents. The simulation results are illustrated in Fig. 3(b). For KD's method, the "no constraint" situation means that the output is selected randomly for comparison, while the "with constraint" randomly select a pitch within the semitone range of ground truth. There is a great gap between the two situations, indicating that there are considerably false positives. The MEA obtains comparable performance when the frequency deviation is smaller than 15 cents compared to the "with constraint" case of KD method.

The standard deviations of errors with respect to the inharmonicity in the aforementioned situations are illustrated in Table 1.



(a) Either fundamental or harmonic frequency shifted



(b) Missing fundamental frequency situation

Fig. 3. Raw pitch accuracy comparison of MEA and KD's methods.

Only the pitches within half semitone range of the ground truth are used for evaluation, since the falsely estimated pitches distribute diversely, greater than 50 cents, and few incorrect estimations may lead to an unfair evaluation of methods. As shown in Table 1, both methods achieve smaller deviations than the inharmonicity of the partials. They perform better in the missing fundamental frequency case, when compared with the other two, indicating that the estimated pitches from higher harmonic partials are more accurate. Moreover, the MEA achieves smaller standard deviations than KD's method in all of these cases, and its degradation is slower with the increase of inharmonicity.

The effect of ς on the virtual pitch estimation is also evaluated. The fundamental frequency is 100 Hz. There are 5 groups of possible harmonic combinations considered, i.e., fundamental-second, fundamental-third, fundamental-fourth, second-third, third-fifth harmonics. There are 1000 frequency pairs for each group. For each pair, one frequency is fixed to be the exact multiple times of fundamental frequency, and the other is multiple times of fundamental frequency plus a shifting frequency with deviation ranging from 0 to 20 cents. The average estimation accuracies for all of the groups vs. ς are evaluated.

The results for different frequency deviations are illustrated in Fig. 4. It is observed from Fig. 4 that given ς, the raw pitch accuracy is higher for the case with smaller deviation. If standard deviation is 0, the results are the same as the Euclidean algorithm when ς is smaller than or equal to 0.3. For a certain deviation, the raw pitch accuracy firstly increases with ς, but decreases after ς is greater than 0.3. The results coincide with our analysis. When ς is small, the harmonic determination condition is very strict, therefore the modified Euclidean algorithm will iterate many times until the strict condition is satisfied. Thus, the raw pitch accuracy is low. The harmonic determination condition looses with the increase of ς, then the raw pitch accuracy is also increased. However, if ς is too large, the false positive rate is increased, leading to reduced raw pitch accuracy. $\varsigma = 0.3$ is the key point where raw pitch accuracy curve alters its direction. ς is set to be 0.15 for the other experiments if there is no specific declaration in this paper.

Similarly, the standard deviations of errors with respect to ς and inharmonicity are also evaluated and shown in Table 2. As displayed in Table 2, the standard deviations of errors are robust to ς for different partial deviations. Thus, the pitch estimation precision is more influenced by inharmonicity rather than ς. However, ς is still a very important parameter, since it is closely related to the raw pitch accuracy, as illustrated in Fig. 4.

4.3.2. Performance on real-world music recordings

The pitch estimation accuracy of the proposed MEA algorithm is also evaluated on the ADC2004 dataset and the results are compared with KD's pair-wise pitch estimation approach [9]. The raw pitch accuracy is employed to evaluate the two pitch candidate calculation algorithms. As we see from Fig. 5(a), the raw pitch accuracy is improved as the candidate number increases for both algorithms. When candidate number is less than 4, KD method (with additional criteria and magnitude weighting) gains higher raw pitch accuracy than the proposed method. If the candidate number is more than 5, the accuracy of "w/o additional criteria" method doesn't increase significantly, and the proposed method surpasses both two variations of KD's method.

The performance of the proposed method is also evaluated on MIREX05 train, MIR-1k datasets. The results are illustrated in Fig. 5(b). The raw pitch accuracy for MIREX05 train is smaller due to the fact that there are some artificially synthesized MIDI recordings with pitches smaller than 50 Hz. But the pitch range

Table 1
Standard deviations of errors with respect to inharmonicity (in cents).

	Deviation = 0	Deviation = 5	Deviation = 10	Deviation = 15	Deviation = 20
Harmonic shifted					
KD	0	2.9955	6.1138	9.0007	12.0682
MEA	0	2.1240	4.3084	6.4042	8.6260
Fundamental shifted					
KD	0	3.0437	6.0954	9.0774	11.9318
MEA	0	2.1325	4.2711	6.4415	8.5117
Missing fundamental					
KD	0	2.0115	3.8000	5.8083	6.7946
MEA	0	1.5522	2.3971	2.8530	3.1224

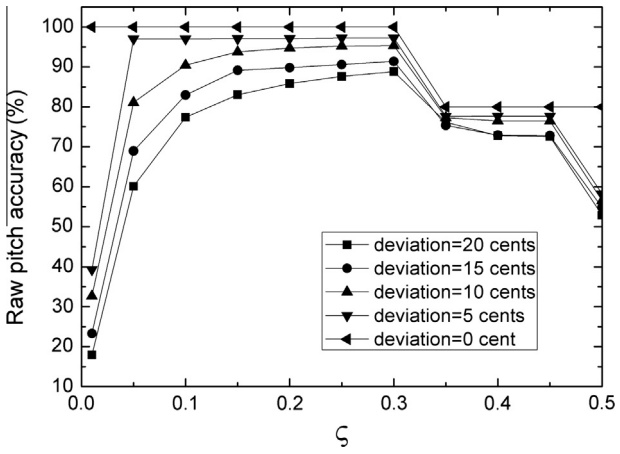


Fig. 4. Raw pitch accuracy of MEA vs. frequency deviation.

of real-world recordings is greater than 50 Hz. It is interesting to see that there is no remarkable improvement if the candidate number is greater than four for MIR-1K. That may originate from the salience function based on the amplitudes of the frame-wise peak pairs. There are two channels for each MIR-1K excerpt. The evaluation is based on 0 dB mixing of melody and accompaniment, indicating that the average power of them are the same. However, the melody is sometimes masked by the percussion or other accompaniments. So the accuracy doesn't increase significantly with the candidate number.

4.4. Performance improvement due to harmonic complement

The frame-wise pitch selection is based on the salience function represented by Eq. (9), which relies on the amplitudes of two sinusoidal components. Generally speaking, the components in lower frequency bands have comparatively higher amplitudes. However, some harmonics (especially the fundamental frequency) are often distorted by the accompaniment or masked by some special singing skills. There are also various combinations of peak pairs. So the pitches with octave errors are often estimated and selected according to the salience function. In the proposed method, these gaps are complemented by the second harmonic and subharmonic pitches. According to our statistical analysis, the average raw pitch accuracy for all of the datasets is improved about 3% due to the harmonic complement.

4.5. Melody extraction based on modified Euclidean algorithm

The proposed main melody extraction method is evaluated on ADC2004, MIREX05 train, MIR-1K and other MIREX melody extraction evaluation datasets. For all of the datasets, only one candidate is selected for each frame. The candidates are grouped and tracked according to the tracking scheme presented in Section 3.2. The raw pitch and raw chroma accuracies are provided in Table 3. As shown in Table 3, the proposed melody extraction method based on the modified Euclidean algorithm can extract the melody in an efficient way. However, the raw chroma accuracies are much higher than the raw pitch accuracies, indicating that there is still much work to do to eliminate the octave errors.

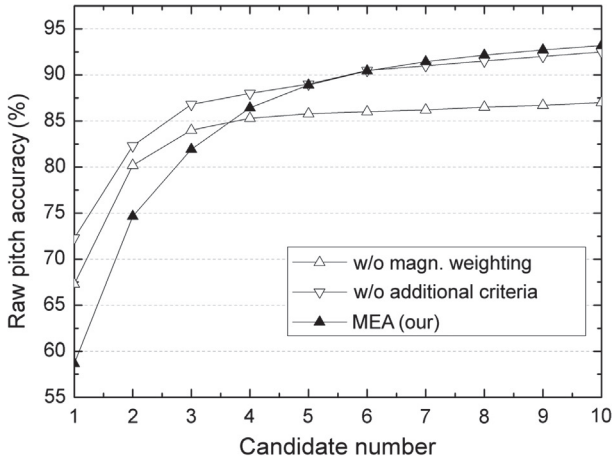
The raw pitch accuracy of the proposed melody extraction algorithm on ADC2004 and MIR-1K is compared with Hsu's normalized subharmonic summation and several variations [28], Cao' harmonic tracking method [16], Dressler's melody extraction based on pair-wise peak evaluation followed by tone rating based tracking [17] and Arora's harmonic cluster tracking algorithm [12]. These methods were reviewed in Section 1. The raw pitch accuracy comparison is shown in Fig. 6. For the proposed method, there is only one candidate, estimated by the proposed modified Euclidean algorithm, considered for each frame. Dressler and Cao's methods are the two best performing algorithms on ADC2004 dataset, however, they do not outperform others any more on the MIR-1K. This fact indicates that their methods don't have advantages for the singing melody extraction. NSHS-DP, Instrument partial deletion +DP and Instrument partial deletion+NSHS-DP proposed by Hsu et al. are designed for singing pitch estimation, so they perform very well on the MIR-1K. The raw pitch accuracy of the proposed method is robust for these two datasets and achieves almost as much raw pitch accuracy as Hsu's methods. The proposed method also outperforms Arora's method for both datasets.

The raw pitch accuracy of melody extraction is about 4% higher than the raw pitch accuracy, when only one candidate is considered as the pitch of each frame without melody contour tracking described in Section 3.2.

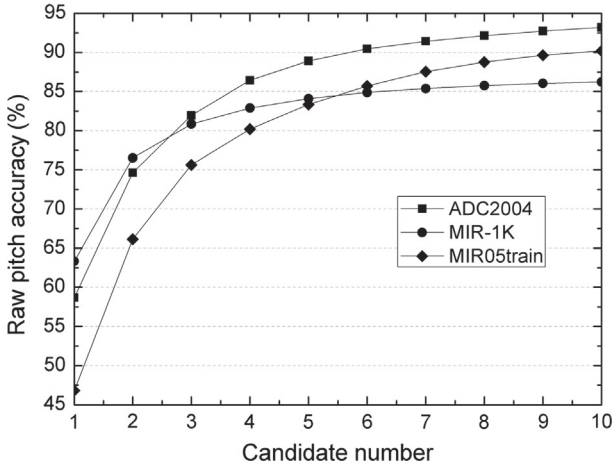
As an example, Fig. 7 illustrates the main melody extraction from one excerpt, where the spectrogram is shown in the top panel, the most salient frame-wise pitches are displayed in the middle panel, and the final main melody extraction output is shown in the bottom panel. From Fig. 7, it's obvious that when melody is dominant, the salient pitches are concentrated around the ground truth, e.g. from 6 s to 10.45 s. Among those intervals where there is no main melody, the estimated most salient pitches are diverse, such as the interval from the beginning to 0.8 s. Moreover, there are still some higher octave errors, e.g. from 5.4 s to 6 s. It is also observed from Fig. 7 that the proposed method works quite well on this excerpt, and can track the dynamic variation of main melody. Unfortunately, there are still false positives from 4.5 s to around 5.3 s. However, when we look back to the spectrogram, there is indeed intense interference during this interval. Hence, it can be concluded that the modified Euclidean algorithm followed by a simple melody tracking scheme can extract melody from polyphonic music. As only one candidate is considered in the aforementioned results, the glass ceiling accuracy can be improved while considering more candidates.

Table 2
Standard deviations of errors with respect to ς and inharmnocity (in cents).

ς	Deviation = 0	Deviation = 5	Deviation = 10	Deviation = 15	Deviation = 20
0.05	0	2.2433	4.3263	5.3564	5.8896
0.1	0	2.2281	4.4326	6.7059	8.4092
0.15	0	2.2358	4.4103	6.7849	8.8417
0.20	0	2.1799	4.4963	6.7626	9.1281
0.25	0	2.2394	4.4496	6.7318	9.1049
0.30	0	2.2049	4.5229	6.7521	9.1042
0.35	0	2.2457	4.4241	6.6998	8.9509
0.40	0	2.2255	4.5592	6.7529	8.8787
0.45	0	2.2593	4.4823	6.8602	9.0231
0.50	0	2.2103	4.4986	6.5766	8.8143



(a) Pitch estimation comparison with KD on ADC2004



(b) Raw pitch accuracy with MEA

Fig. 5. Raw pitch estimation of the MEA on real-world recordings.

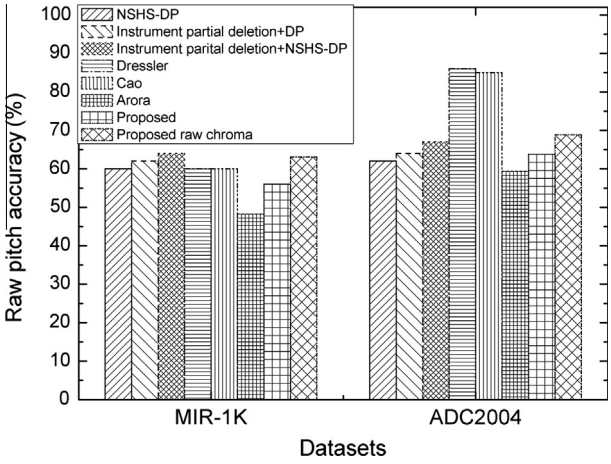


Fig. 6. Performance comparison for singing pitch extraction.

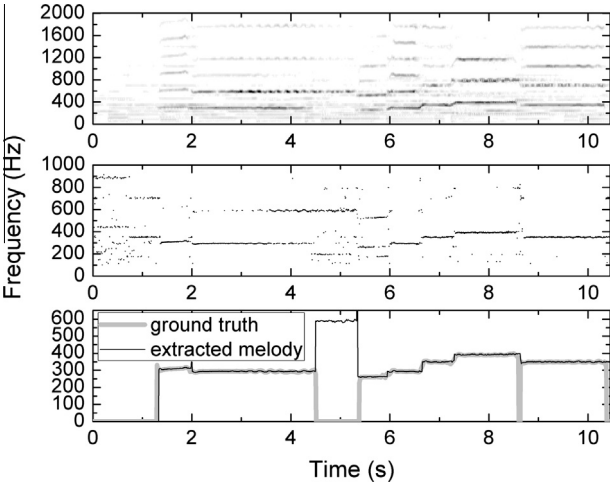


Fig. 7. Melody extraction from one excerpt.

Table 3
Melody extraction accuracy of three datasets.

Dataset	Raw pitch accuracy (%)	Raw chroma accuracy (%)
ADC2004	63.82	68.85
MIREX05 train	64.08	66.25
MIR-1K	56.07	63.09
ISMIR04	68.14	73.41
INDIAN08	57.59	67.02
MIREX09 0 dB	64.72	72.939

5. Conclusions

In this paper, a main melody extraction method based on the modified Euclidean algorithm is proposed. Specifically, the Euclidean algorithm for positive integers is first generalized to float numbers and used for frame-wise pitch candidate estimation. Next, the salience function of each pitch candidate is constructed relying on the amplitudes of the two corresponding spectral peaks. Afterwards, melodic contours are formed according to the continuity principle by the candidates. Only the estimated pitch candidates are delivered for melody contour construction and selection, reducing memory requirement. Few candidates are used for melody contour tracking, which makes the tracking stage very simple. Finally, the proposed method is evaluated on some melody extraction datasets. The experimental results show that the proposed method requires few pitch candidates to achieve good main melody extraction performance.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 61172107, 61172110), National High Technology Research and Development Program (863 Program) of China (No. 2015AA016306), Major Projects in Liaoning Province Science and Technology Innovation, China (No. 201302001), Natural Science Foundation of Liaoning Province, China (No. 2013020018), General Project of Education Department of Liaoning Province, China (No. L2015135), and Fundamental Research Funds for the Central Universities of China (Nos. DUT13LAB06, DC201502060404).

The authors would like to thank the editors and anonymous reviewers for their valuable comments.

References

[1] Salamon J, Serra J, Gmez E. Tonal representations for music retrieval: from version identification to query-by-humming. Int J Multimedia Inform Ret 2013;2:45–58. http://dx.doi.org/10.1007/s13735-012-0026-0.

[2] Wang C-C, Jang J-SR, Wang W. An improved query by singing/humming system using melody and lyrics information. In: The 11th international society for music information retrieval conference (ISMIR), Florida, USA. p. 45–50.

[3] Foucard R, Durrieu J-L, Lagrange M, Richard G. Multimodal similarity between musical streams for cover version detection. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). Dallas (USA): IEEE; 2010. p. 5514–7. http://dx.doi.org/10.1109/ICASSP.2010.5495217.
[4] Ryyänen MP, Klapuri AP. Automatic transcription of melody, bass line, and chords in polyphonic music. Comput Music J 2008;32:72–86. http://dx.doi.org/10.1162/comj.2008.32.3.72.
[5] Otsuka T, Nakadai K, Takahashi T, Ogata T, Okuno HG. Real-time audio-to-score alignment using particle filter for coplayer music robots. EURASIP J Appl Signal Process 2011;2011:1–13. http://dx.doi.org/10.1155/2011/384651.
[6] Goto M. A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. Speech Commun 2004;43:311–29. http://dx.doi.org/10.1016/j.specom.2004.07.001.
[7] Salamon J, Gmez E. Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Trans Audio Speech Lang Process 2012;20:1759–70. http://dx.doi.org/10.1109/TASL.2012.2188515.
[8] Paiva RP, Mendes T, Cardoso A. Melody detection in polyphonic musical signals: exploiting perceptual rules, note salience, and melodic smoothness. Comput Music J 2006;30:80–98. http://dx.doi.org/10.1162/comj.2006.30.4.80.
[9] Dressler K. Pitch estimation by the pair-wise evaluation of spectral peaks. In: AES 42nd international conference. Ilmenau, Germany: Audio Engineering Society; 2011. p. 1–10. http://www.aes.org/e-lib/browse.cfm?elib=15960.
[10] Durrieu J-L, Richard G, David B, Fvotte C. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. IEEE Trans Audio Speech Lang Process 2010;18:564–75. http://dx.doi.org/10.1109/TASL.2010.2041114.
[11] Hsu C-L, Wang D, Jang J-S. A trend estimation algorithm for singing pitch detection in musical recordings. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). Prague (Czech Republic): IEEE; 2011. p. 393–6. http://dx.doi.org/10.1109/ICASSP.2011.5946423.
[12] Arora V, Behera L. On-line melody extraction from polyphonic audio using harmonic cluster tracking. IEEE Trans Audio Speech Lang Process 2013;21:520–30. http://dx.doi.org/10.1109/TASL.2012.2227731.
[13] Rao V, Rao P. Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. IEEE Trans Audio Speech Lang Process 2010;18:2145–54. http://dx.doi.org/10.1109/TASL.2010.2042124.
[14] Ikemiya Y, Yoshii K, Itoyama K. MIREX2014: audio melody extraction. In: Music information retrieval evaluation eXchange (MIREX), vol. 15; 2014. p. 1–2.
[15] Degani A, Leonardi R, Migliorati P, Peeters G. A pitch salience function derived from harmonic frequency deviations for polyphonic music analysis. In: 17th International conference on digital audio effects, Erlangen, Germany. https://hal.archives-ouvertes.fr/hal-01254091/.
[16] Cao C, Li M, Liu J, Yan Y. Singing melody extraction in polyphonic music by harmonic tracking. In: The 8th international society for music information retrieval conference, Vienna, Australia. p. 373–4.
[17] Dressler K. An auditory streaming approach for melody extraction from polyphonic music. In: The 12th international society for music information retrieval conference (ISMIR2011), Florida, USA. p. 19–24.
[18] Salamon J, Gomez E, Ellis DP, Richard G. Melody extraction from polyphonic music signals: approaches, applications, and challenges. IEEE Signal Process Mag 2014;31:118–34. http://dx.doi.org/10.1109/MSP.2013.2271648.
[19] Keiler F, Marchand S. Survey on extraction of sinusoids in stationary sounds. In: The 5th international conference on digital audio effects (DAFx), Hamburg, Germany. p. 51–8. https://hal.archives-ouvertes.fr/hal-00308216/.
[20] Abe T, Kobayashi T, Imai S. The IF spectrogram: a new spectral representation. In: International symposium on simulation, visualization and auralization for acoustic research and education, Tokyo, Japan. p. 423–30. http://ci.nii.ac.jp/naid/10007458656/.
[21] Charpentier F. Pitch detection using the short-term phase spectrum. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP). Tokyo (Japan): IEEE; 1986. p. 113–6. http://dx.doi.org/10.1109/ICASSP.1986.1169123.
[22] Cheveigne AD. Pitch perception models. UK: Oxford University Press; 2005. http://dx.doi.org/10.1109/ICASSP.1986.1168883.
[23] Howard DM, Angus JAS. Acoustics and psychoacoustics. Massachusetts (USA): Focal Press; 2009.
[24] Bradley GH. Algorithm and bound for the greatest common divisor of n integers. Commun ACM 1970;13:433–6. http://dx.doi.org/10.1145/362686.362694.
[25] Marolt M. On finding melodic line in audio recordings. In: The 7th international conference on digital audio effects (DAFx). Naples, Italy. p. 1–5.
[26] Bregman AS. Auditory scene analysis: the perceptual organization of sound. America: MIT Press; 1994.
[27] MIREX homepage. <http://www.music-ir.org/mirex/wiki/MIREX_HOME>.
[28] Hsu C-L, Chen L-Y, Jang J-SR, Li H-J. Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement. In: The 10th international society for music information retrieval conference. Kobe, Japan: Citeseer; 2009. p. 201–6.