Automatic Singing Performance Evaluation Using Accompanied Vocals as Reference Bases[*]

WEI-HO TSAI, CIN-HAO MA AND YI-PO HSU
Department of Electronic Engineering
National Taipei University of Technology
Taipei, 106 Taiwan
E-mail: {whtsai; t101419012}@ntut.edu.tw

This work aims to develop an automatic singing evaluation system for general public. Given a CD/mp3 song recording as the reference basis, the proposed system rates a user's singing performance by comparing it with the vocal in the song recording. This modality allows users to not only enjoy listening to and singing with CD/mp3 songs but also know how well or bad they sing. However, as a majority of songs contain background accompaniments during most or all vocal passages, directly comparing a user's singing performance with the signals in a song recording does not make sense. To tackle this problem, we propose methods to extract pitch-, volume-, and rhythm-based features of the original singer in the accompanied vocals. Our experiment shows that the results of automatic singing evaluation are close to the human rating, where the Pearson product-moment correlation coefficient between them is 0.8. The results are also comparable to those in a previous work using Karaoke music as reference bases, where the latter's task is considered to be easier than that of this work.

Keywords: accompanied vocal, pitch, rhythm, singing evaluation, volume

1. INTRODUCTION

Karaoke is a popular entertainment and practice means for amateur singers, where people sing along to a pre-recorded accompanying music of a selected song while various scenes are displayed on a screen so that it looks like professional artists are performing. Thanks to technological innovations on electronic and communication devices, Karaoke has become ubiquitous, ranging from Karaoke jukebox, Karaoke bar, TV Karaoke on demand, in-car Karaoke, to mobile Karaoke apps. Most of the karaoke systems today come with a number of standard features such as song selection search, key changer, lyric prompt, pitch graph, performance scoring, and more. Relying on these enticing features to learn how to sing better and challenge other people to beat your score, however, may not be a satisfying way to interpret and evaluate your singing performances. The major problem arises from the fact that most existing Karaoke designers do not investigate the required techniques seriously. A vast majority of Karaoke apparatuses use singing energy as a unique cue for performance scoring, while some apparatuses even display a random score for fun. As a result, the presenting scores in the existing Karaoke apparatuses is usually nothing to do with the singing skill and considered useless to users. Thus, there is a high need to develop reliable singing scoring techniques to make this function functional.

Received April 9, 2014; revised August 12 & November 10, 2014; accepted December 2, 2014.
Communicated by Chung-Hsien Wu.

[*] This work was supported in part by the National Science Council under Grants NSC101-2628-E-027-001 and MOST103-2218-E-027-006-MY3.

This research effort focuses on developing an automatic singing evaluation system for general public, but not for professional singers. Although so far there have been several studies [1-13] to this end, most of them are reported in patent documentation, which only describe their implementation details and fail to present the theoretical foundation and qualitative analysis conducted to validate their methods. Only very few studies are reported in scientific literature. The most thorough investigation of this research topic is a work reported in [13]. It comprehensively discusses the strategies and acoustic cues for singing performance evaluation. The strategy depends heavily on the reference basis (or ground truth), which is used to measure the correctness of a singing performance in terms of pitch, volume, rhythm, and so on. Roughly speaking, there are five types of reference basis: (1) music scores and lyrics; (2) symbolic music, *e.g.*, MIDI files; (3) CD/mp3 music; (4) Karaoke VCD music; (5) solo vocal track. Each type of reference basis has its own pros and cons, as summarized in Table 1. Among the five types, CD/mp3 music is the easiest one to acquire for general public, since the others can only be available when a song has been released and become popular for a period. Thus, the proposed singing evaluation system is built on using CD/mp3 music as a reference basis. This differs from the work in [13], which used Karaoke VCD music as the reference basis, and the work in [5], which used MIDI files as the reference basis.

Table 1. Pros and cons of the five types of reference basis for singing performance evaluation.

Reference Basis	Pros	Cons
Music Scores and Lyrics	Easy to Use for System Designers	Relying on Human Processing
Symbolic Music, *e.g.*, MIDI Files	Easy to Use for System Designers	Not Always Available
CD/mp3 Music	Easy to Acquire	Difficult to Handle
Karaoke VCD Music	Easy to Integrate with Some Karaoke Systems	Not Popular
Solo Vocal Track	Easy to Use for System Designers	Difficult to Acquire

In essence, the proposed system rates a user's singing performance by comparing it with the vocal in the CD/mp3 song recording. This modality allows users to not only enjoy listening to and singing with CD/mp3 songs but also know how well or bad they sing. Since the proposed system does not rely on any dedicated audio formats, such as Karaoke VCD music, Digital Video Systems (DVS) or the Laser Disc (LD) karaoke systems, in which accompaniments are stored in separated tracks, it is particularly suitable for mobile apps. More specifically, as long as a user has a regular CD/mp3 song recording, where even the accompaniments and vocals are mixed, singing evaluation can be performed whenever the user sings to our system.

However, as a vast majority of songs contain background accompaniments during most or all vocal passages, directly comparing a user's singing performance with the signals in a song recording does not make sense. To tackle this problem, we propose methods to extract pitch-, volume-, and rhythm-based features of the original singer in the accompanied vocals. This task is more difficult than the one investigated in [13], where the latter can use the accompaniment-only track to help extract vocal information from the accompanied vocal track. Despite the difficulty, our experiment shows that the results of the proposed singing evaluation system are comparable to those of the system in [13] and also close to the human rating. Table 2 summarizes the major contribution of this work, compared to the work in [13].

Table 2. Major contribution of this work, compared to a previous work in [13].

	The Work in [13]	This Work
Reference Basis for Singing Performance Evaluation	Karaoke VCD Music, Encompassing Two Distinct Channels: (1) the Accompaniment only; (2) a Mixture of the Lead Vocals and Background Accompaniment	CD/mp3 Music, Consisting of Two Similar Accompanied Vocal Channels
Technical Features	Using the Accompaniment-only Track to Help Extract Vocal Information from the Accompanied Vocal Track	Extracting Vocal Information from the Accompanied Vocal Track without Using Any Other Audio Resources
Application Niche	Karaoke Apparatuses	Mobile Devices
Other Traits	First Study of Integrating Pitch, Volume, and Rhythm Features for Singing Performance Evaluation	Proposing a Simple-yet-effective Rhythm-based Rating Method

The remainder of this paper is organized as follows. Section 2 presents the methodology of the proposed singing evaluation system. Section 3 discusses our experiment results. In Section 4, we present our conclusions and indicate the directions of our future work.

2. METHODOLOGY

When a singing piece is evaluated, the proposed system performs volume-based rating, pitch-based rating, and rhythm-based rating, using the specified song (accompanied vocal recording) extracted from CD/mp3 music as a reference basis. Similar to the strategy used in [13], the resulting scores from each component are then combined using a weighted sum method:

$$\text{overall score} = \sum_{i=1}^{3} w_i S_i, \tag{1}$$

where S_1, S_2, and S_3 are the scores obtained with pitch-based rating, volume-based rating, and rhythm-based rating, respectively; w_1, w_2, and w_3 are the adjustable weights that sum to 1. However, since the reference bases are the accompanied vocal recordings extracted from CD/mp3 music rather than the Karaoke VCD music considered in [13], the ways to exploiting the pitch-, volume-, and rhythm-based features in the recordings must be specifically tailored to handle the interference arising from the background accompaniments.

2.1 Pitch-based Rating

Pitch represents the degree of highness or lowness of a tone. In singing, pitch is related to the notes performed by a singer. To sing in tune, a prerequisite is to perform a sequence of correct notes, each with appropriate duration. By representing musical notes as MIDI numbers, we can compute the difference between a sequence of notes sung in an evaluated recording with the one sung in the reference recording.

As shown in Fig. 1, the pitch-based rating starts by converting the waveform of a singing recording into a sequence of MIDI notes **o** = {o_1, o_2,…, o_T}. Our method is similar to that in [14], which consists of the following steps.

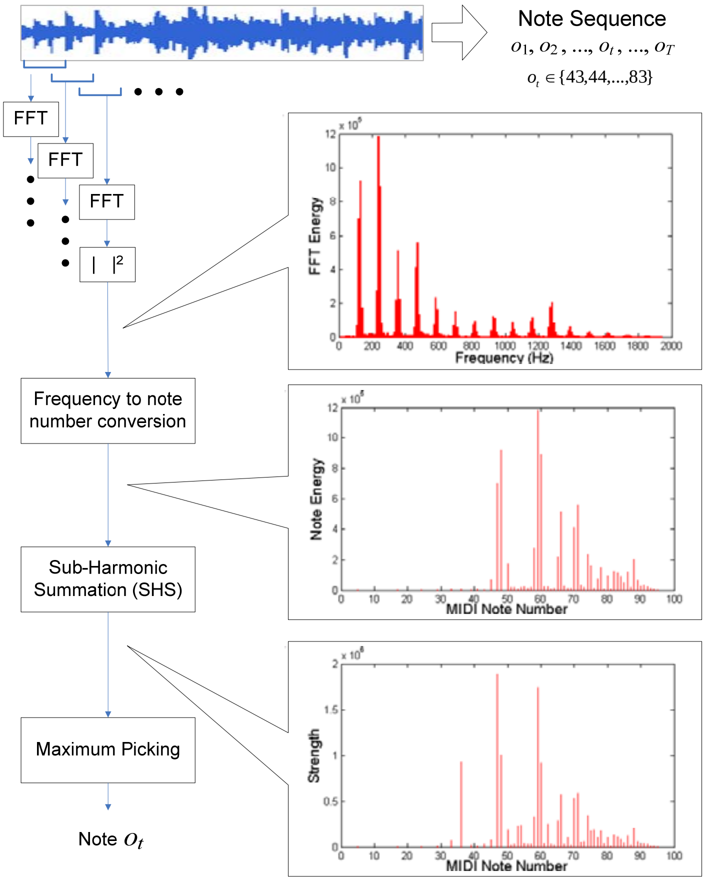


Fig. 1. Conversion of a waveform recording into a MIDI note sequence.

1) Dividing the waveform signal into frames using a sliding Hamming window.
2) Performing Fast Fourier Transform (FFT) with respect to each frame.
3) Computing the signal's energy with respect to each FFT index (frequency bin) in a frame
4) Estimating the signal’s energy with respect to each MIDI note number in a frame according to the conversion of Hz to MIDI note:

$$\text{MIDI note} = \left\lfloor 12 \cdot \log_2\left(\frac{\text{Hz}}{440}\right) + 69.5 \right\rfloor, \tag{2}$$

where $\lfloor\ \rfloor$ is a floor operator.

5) Summing the signal's energy belonging to a note and its harmonic note numbers to obtain a strength value, *i.e.*, the strength of the *m*th note in the *t*th frame is obtained by

$$y_{t,m} = \sum_{c=0}^{C} h^c \, e_{t,m+12c} \,, \tag{3}$$

where $e_{t,m}$ is the signal's energy belonging to the *m*th note in the *t*th frame, *C* is the number of harmonics considered, and *h* is a positive value less than 1 that discounts the contribution of higher harmonics.

6) Determining the sung note in the *t*th frame by choosing the note number associated with the largest value of the strength accumulated for adjacent ±*B* frames, *i.e.*,

$$o_t = \arg\max_{1 \le m \le M} \sum_{b=-B}^{B} y_{t+b,m} \,, \tag{4}$$

where *M* is the number of the possible notes performed by a singer.

7) Removing jitters between adjacent frames by replacing each note with the local median of notes of its neighboring ±*B* frames.

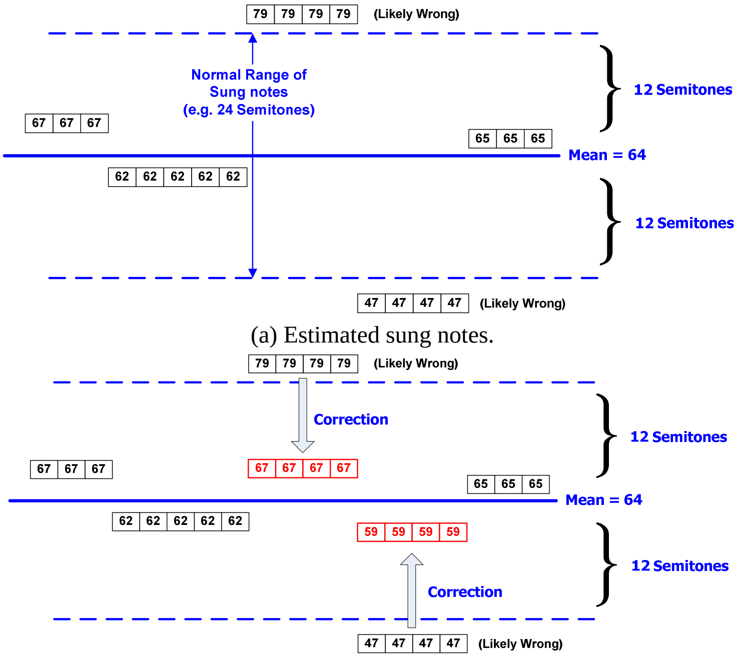
However, the above method is only suitable for extracting the note sequence of a singing recording with no background accompaniment. Since there is always background accompaniment in most of the vocal passages in popular music, the note number associated with the largest value of the strength may not be produced by the singer, but the instrumental accompaniment instead. To solve this problem, we propose a method to correct the error estimation of sung notes.

The basic strategy of our method is to identify abnormal elements in a note sequence and forces them back to the normal notes. The abnormality in a note sequence generally comes in two types of errors: short-term errors and long-term errors. Short-term errors refer to rapid changes (*e.g.*, jitters) between adjacent frames. This type of error can be corrected by using median filtering, which replaces each note with the local median of the notes of its neighboring frames. Long-term errors, on the other hand, refer to a succession of estimated notes that are not produced by a singer. Our experiments found that such wrong notes are often several octaves above or below the true sung notes, which mainly arise from the background accompaniment. This might be because the background accompaniment often contains notes several octaves above or below those of the singing so that the mixture of the lead vocals and the background accompaniment is harmonic. As a consequence, long-term errors often result in the range of the estimated notes in a sequence being wider than that of the true sung note sequence.

According to our statistics on pop music, the sung notes in a verse or chorus section seldom vary by more than 24 semitones. Thus, if it is found that the range of the estimated notes in a sequence is wider than the normal range, we can adjust the suspect notes by shifting them several octaves up or down, so that the range of the notes in an adjusted sequence conforms to the normal range. Specifically, let $\mathbf{o} = \{o_1, o_2, \ldots, o_T\}$ denote a note sequence estimated using Eq. (4). The adjusted note sequence $\mathbf{o}' = \{o'_1, o'_2, \ldots, o'_T\}$ is obtained by

$$o'_t = \begin{cases} o_t \quad , & \text{if } |o_t - \bar{o}| \le (Z/2) \\ o_t - 12 \times \left\lfloor \dfrac{o_t - \bar{o} + Z/2}{12} \right\rfloor, & \text{if } o_t - \bar{o} > (Z/2) \;, \\ o_t - 12 \times \left\lfloor \dfrac{o_t - \bar{o} - Z/2}{12} \right\rfloor, & \text{if } o_t - \bar{o} < (-Z/2) \end{cases} \tag{5}$$

where Z is the normal range of the sung notes in a sequence, *e.g.*, $Z = 24$, and $\bar{o}$ is the mean note computed by averaging all the notes in **o**. In Eq. (5), a note o_t is deemed a wrong note that must be adjusted if it is too far from $\bar{o}$, *i.e.*, $|o_t - \bar{o}| > Z/2$. The adjustment is done by shifting the wrong note $\lfloor(o_t - \bar{o} + Z/2)/12\rfloor$ or $\lfloor(o_t - \bar{o} - Z/2)/12\rfloor$ octaves.



(a) Estimated sung notes.

(b) Modification of the notes in (a) using Eq. (5).

Fig. 2. Example of the long term correction.

Fig. 2 shows an example of the long term correction. In Fig. 2 (a), the estimated sung note sequence is {67,67,67,62,62,62,62,62,79,79,79,79,65,65,65} and its mean is 64. If we consider the normal range of sung notes is 24semitones (±12 semitones), then notes {79,79,79,79} are likely incorrect, because they are 15 semitones above the mean (64), which exceed the normal range. Similarly, notes {47,47,47,47} are likely incorrect, because they are 17 semitones below the mean (64), which exceed the normal range as well. In Fig. 2 (b), notes {79,79,79,79} and {47,47,47,47} are modified by {67,67,67,67} and {59,59,59,59}, respectively, using Eq. (5).

With the note sequences $\mathbf{O} = \{o_1, o_2, ..., o_T\}$ and $\mathbf{O}' = \{o'_1, o'_2, ..., o'_{T'}\}$ computed from the reference recording and an evaluated singing recording, respectively, the pitch-based rating can be done by comparing the difference between **O** and **O**′. However, since the lengths of the two sequences are usually different, computing their Euclidean distance directly is infeasible. To deal with this problem, we apply Dynamic Time Warping (DTW) to find the temporal mapping between **O** and **O**′.

DTW begins by constructing a distance matrix $\mathbf{D} = [D(t, t')]_{T \times T'}$, where $D(t, t')$ is the distance between note sequences $\{o_1, o_2, ..., o_t\}$ and $\{o'_1, o'_2, ..., o'_{t'}\}$, computed using:

$$D(t,t') = \min \begin{cases} D(t-2,t'-1) + 2 \times d(t,t') \\ D(t-1,t'-1) + d(t,t') - \rho \\ D(t-1,t'-2) + d(t,t') \end{cases}, \tag{6}$$

and

$$d(t, t') = |o_t - o_{t'}|, \tag{7}$$

where ρ is a small constant that favors the mapping between notes o_t and $o_{t'}$, given the distance between note sequences $\{o_1, o_2, ..., o_{t-1}\}$ and $\{o'_1, o'_2, ..., o'_{t'-1}\}$. The boundary conditions for the above recursion are defined by

$$\begin{cases} D(1,1) = d(1,1) \\ D(t,1) = \infty,\ 2 \le t \le T \\ D(1,t') = \infty,\ 2 \le t' \le T' \\ D(2,2) = d(1,1) + d(2,2) - \rho \\ D(2,3) = d(1,1) + d(2,2) \\ D(3,2) = d(1,1) + 2 \times d(2,2) \\ D(t,2) = \infty,\ 4 \le t \le T \\ D(2,t') = \infty,\ 4 \le t' \le T' \end{cases}. \tag{8}$$

After the distance matrix **D** is constructed, the DTW distance between **O** and **O**′ can be evaluated by

$$\mathrm{DTWDist}(\mathbf{O}, \mathbf{O}') = \begin{cases} \min\limits_{T/2 \le t' \le \min(2T, T')} [D(T,t')/T], & \text{if } \frac{T}{2} \le T' \le 2T \\ \infty & \text{, otherwise} \end{cases}, \tag{9}$$

where we assume that the length of a test singing should be no shorter than a half length of the reference singing and no longer than a double length of the reference singing. The distance DTWDist(**O**,**O**′) is then converted to a pitch-based score between 0 and 100:

$$S_1 = 100 \cdot k_1 \exp[k_2 \cdot \mathrm{DTWDist}(\mathbf{O}, \mathbf{O}')], \tag{10}$$

where k_1 and k_2 are tunable parameters used to control the distribution of S_1.

2.2 Volume-based Rating

Our basic strategy for volume-based rating is to represent an evaluated singing signal and the reference singing signal as short-term energy sequences, and then compare the difference between the two sequences. However, as the reference singing signal is intermixed with background accompaniment, it is impossible to acquire the reference singing signal's short-term energy sequence directly from the CD music data. To solve this problem, we use the sung note correction method described in Section 2.1 to help estimate the reference singing signal's energy. Specifically, after the reference recording is converted from its waveform representation into a note sequence $\mathbf{O} = \{o_1, o_2, ..., o_T\}$, with the short-term and long-term note correction being performed, the short-term energy sequence $\mathbf{G} = \{g_1, g_2, ..., g_T\}$ is obtained using

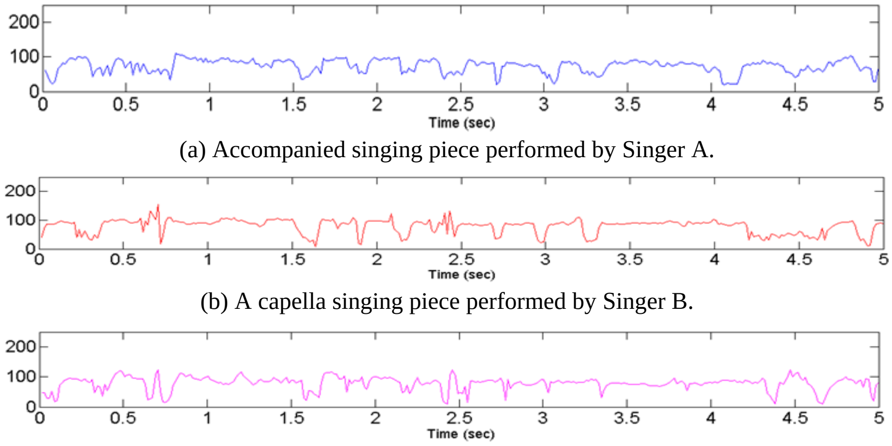
$$g_t = e_{t,o_t}, 1 \le t \le T \tag{11}$$

which is the energy of note o_t in the tth frame.

Given an evaluated singing recording, we compute its short-term energy sequence, $\mathbf{G}'$, and apply the DTW to measure distance, DTWDist($\mathbf{G}$,$\mathbf{G}'$), between $\mathbf{G}$ and $\mathbf{G}'$. Then, a volume-based score is obtained using

$$S_2 = 100 \cdot q_1 \exp[q_2 \cdot \text{DTWDist}(\mathbf{G}, \mathbf{G}')], \tag{12}$$

where q_1 and q_2 are tunable parameters used to control the distribution of S_2. Fig. 3 shows an example of short-term energy sequences, respectively, computed from an accompanied singing piece and two *a capella* singing pieces, in which all the three sequences are associated with the same song but different singers. We can see that the contours of the short-term energy sequences in Figs. 3 (a), (b), and (c) are similar.

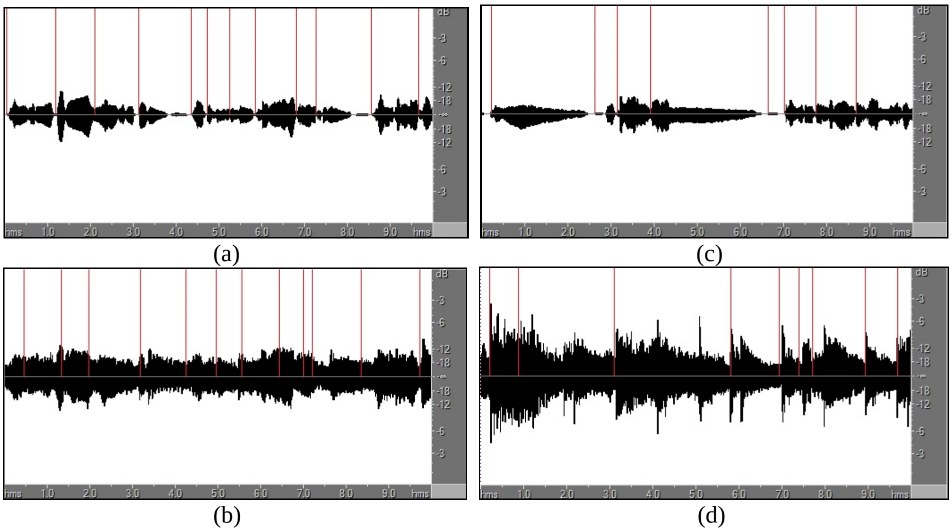


(a) Accompanied singing piece performed by Singer A.

(b) A capella singing piece performed by Singer B.

(c) A capella singing piece performed by Singer C.

Fig. 3. Example of short-term energy sequences from three different singers singing the same song.



(a) (c)

(b) (d)

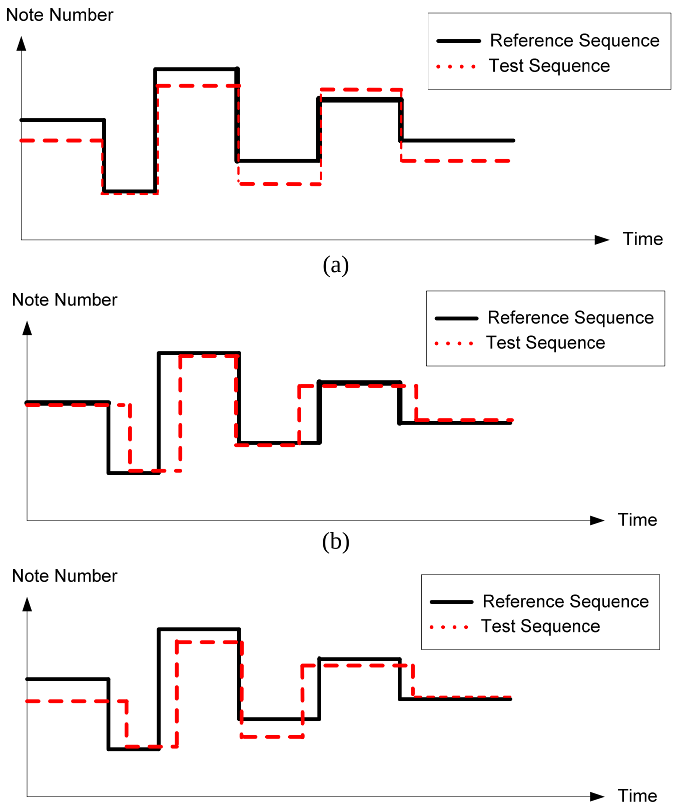
Fig. 4. Examples of the note detection with *SuperFlux*, in which (a) is a solo singing clip; and (b) is an accompanied singing clip by manually mixing (a) with the corresponding Karaoke accompaniment; (c) is a solo singing clip other than (a); and (d) is an accompanied singing clip by manually mixing (b) with the corresponding Karaoke accompaniment. The vertical straight lines represent the detected note onsets.

2.3 Rhythm-based Rating

Rhythm is related to the onset and duration of successive notes and rest performed by a singer. Thus, an intuitive approach to rhythm-based rating is to detect and compare the onsets of notes sung in the reference recording and an evaluated singing recording. There are a number of note onset detection algorithms [16] available to apply here, with *SuperFlux* [17] being the current state of the art. However, all the existing algorithms are designed for the pure vocal or pure instrumental music, and hence they may not work well for detecting the note onsets of the vocals accompanied with background music. Fig. 4 shows some examples of the note detection with *SuperFlux*. We can see from Fig. 4 that the detected onsets marked with the vertical lines are significantly different in between a pure vocal signal and its accompanied version. As the reference recordings in our task are accompanied vocals, it is expected that the detected note onsets[1] cannot reliably used for rhythm-based rating.

Instead of locating note onsets in a singing recording, we propose a rhythm-based rating method by exploiting the information from the note sequence used in the pitch-based rating. Fig. 5 shows an example of simulated note sequence for ease of discussion. In Fig. 5 (a), we can see that the test singing recording is of correct rhythm but wrong pitch, compared to the reference singing recording. On the contrary, we can see from Fig. 5 (b) that the test singing recording is of correct pitch but wrong rhythm. In Fig. 5 (c), it is clear that there are errors in both rhythm and pitch of the test singing recording. Accordingly, the rhythm-based rating may be done by measuring and subtracting the pitch-related errors from the total errors in a test singing recording note sequence.

[1] Using dataset DB-1 described in Sec. 3.1, the recall, precision, and F-measure obtained with SuperFlux were 57.3%, 38.2%, and 45.8% respectively, based on an error tolerance of 100ms.



(a)

(b)

(c)

Fig. 5. (a) Errors in pitch; (b) Errors in rhythm; (c) Errors in both pitch and rhythm.

Let $\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$ and $\mathbf{O}' = \{o'_1, o'_2, \ldots, o'_{T'}\}$ be the note sequences extracted from the reference recording and a test singing recording, respectively. We can observe the following four cases.

(i) If **O** and **O′** are consistent in both pitch and rhythm, then obviously both the Euclidean distance and DTW distance between **O** and **O′** are zero, *i.e.*, EucDist(**O**,**O′**) = DTWDist(**O**,**O′**) = 0.
(ii) If **O** and **O′** are consistent in pitch but inconsistent in rhythm, then EucDist(**O**,**O′**) > DTWDist(**O**,**O′**) = 0, because DTW can absorbs the difference of rhythm between **O** and **O′**.
(iii) If **O** and **O′** are inconsistent in pitch but consistent in rhythm, then EucDist(**O**,**O′**) = DTWDist(**O**,**O′**) > 0.
(iv) If **O** and **O′** are inconsistent in both pitch and rhythm, then EucDist(**O**,**O′**) > DTWDist(**O**,**O′**) > 0.

Thus, the errors in rhythm can be characterized by EucDist(**O**,**O′**) – DTWDist(**O**,**O′**). For rhythm-based rating, we convert the errors into a rhythm-based score between 0 and 100:

$$S_3 = 100 \cdot r_1 \exp\{r_2 \cdot [\mathrm{EucDist}(\mathbf{O},\mathbf{O}') - \mathrm{DTWDist}(\mathbf{O},\mathbf{O}')]\}, \tag{13}$$

where r_1 and r_2 are tunable parameters used to control the distribution of S_3. Fig. 6 summaries the overall procedure of the proposed singing-evaluation system.

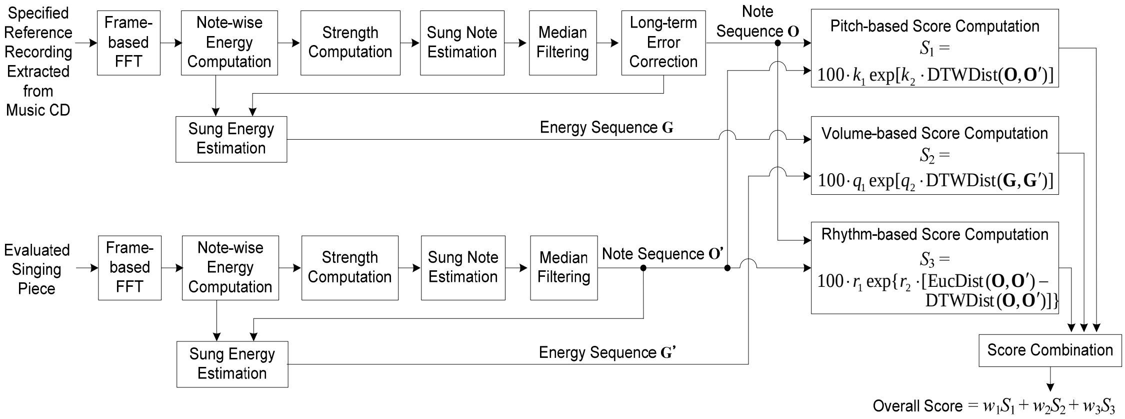

Specified Reference Recording Extracted from Music CD
Frame-based FFT
Note-wise Energy Computation
Strength Computation
Sung Note Estimation
Median Filtering
Long-term Error Correction
Note Sequence $\mathbf{O}$
Pitch-based Score Computation
$S_1 = 100 \cdot k_1 \exp[k_2 \cdot \mathrm{DTWDist}(\mathbf{O},\mathbf{O}')]$
Sung Energy Estimation
Energy Sequence $\mathbf{G}$
Volume-based Score Computation
$S_2 = 100 \cdot q_1 \exp[q_2 \cdot \mathrm{DTWDist}(\mathbf{G},\mathbf{G}')]$
Rhythm-based Score Computation
$S_3 = 100 \cdot r_1 \exp\{r_2 \cdot [\mathrm{EucDist}(\mathbf{O},\mathbf{O}') - \mathrm{DTWDist}(\mathbf{O},\mathbf{O}')]\}$
Evaluated Singing Piece
Frame-based FFT
Note-wise Energy Computation
Strength Computation
Sung Note Estimation
Median Filtering
Note Sequence $\mathbf{O}'$
Sung Energy Estimation
Energy Sequence $\mathbf{G}'$
Score Combination
Overall Score $= w_1S_1 + w_2S_2 + w_3S_3$

Fig. 6. The overall procedure of the proposed singing-evaluation system.

3. EXPERIMENTS

3.1 Music Database

Two music datasets were created by ourselves. The first one, denoted by DB-1, contains 20 Mandarin song clips extracted from music CDs. For computational efficiency, each extracted music track was downsampled from 44.1 kHz to 22.05 kHz and stored as PCM wave. Each clip contains a verse or chorus part of song, which ranges in duration from 25 to 40 seconds. The second dataset, denoted by DB-2, has been created and used in [13]. It contains singing samples recorded by in a quiet room. We employed 25 singers to record for solo vocal parts of the 20 Mandarin song clips. Every singer performed solely with a Karaoke machine, which sang along with onscreen guidance to popular song recordings from which the vocals have been removed. The Karaoke accompaniments were output to singer's headset and were not captured in the recordings. The recordings were stored in mono PCM wave with 22.05 kHz sampling rate and 16-bit quantization level.

As described in [13], 10 among the 25 singers in DB-2, marked by Group I, are considered to have good singing capabilities. The other 10 among the 25 singers are those who like to sing Karaoke, but their singing capabilities are far from professional. They are marked by Group II. The remaining 5 among the 25 singers, marked by Group III, are considered to have poor singing capabilities. They sometimes cannot follow the tune, and some of them even never sing Karaoke before. To establish the ground truth for automatic singing evaluation, the singing recordings were rated independently by four musicians we employed. The ratings were done in terms of technical accuracy in pitch, volume, rhythm, and combination thereof, in which the rating results given by the four musicians were averaged to form a reference score for each singing recording.

We further divided Dataset DB-2 into two subsets. The first subset, denoted by DB-2A, contains 150 recordings performed by 10 singers, in which 2 singers were selected from Group I, the other 6 from Group II, and the remaining 2 from Group III. The second subset, denoted by DB-2B, contains the remaining recordings of DB-2 not covered in DB-2A. We used DB-2B to tune the parameters in Eqs. (1), (10), (12) and (13), and used DB-2A to test our system. Table 3 summarizes the datasets used in this paper.

Table 3. The dataset used in this paper.

Dataset	Content	Purpose
DB-1	20 Mandarin song clips extracted from music CDs	Reference bases
DB-2-A	Mandarin singing a capella clips performed by 10 amateur singers; each 15 song clips	System evaluation
DB-2-B	Mandarin singing a capella clips performed by 15 amateur singers; each 20 song clips. The singers and songs are different from those in DB-2-A	System parameter tuning

3.2 Experiment Results

3.2.1 Experiments on pitch-based rating

First, we examined the validity of our method for converting waveform recordings into MIDI note sequences. All the recordings in DB-1 and DB-2-A were manually annotated with the groundtruth MIDI note sequences. In our system, we set the length of frame, FFT size, parameters *C*, *h*, and *B*, in Eqs. (3) and (4) to be 30-ms and 2048, 2, 0.8, and 2, respectively. The performance of the conversion was characterized by the frame accuracy:

$$\text{Accuracy (\%)} = \frac{\text{No. of correctly converted frames}}{\text{No. of total frames}}.$$

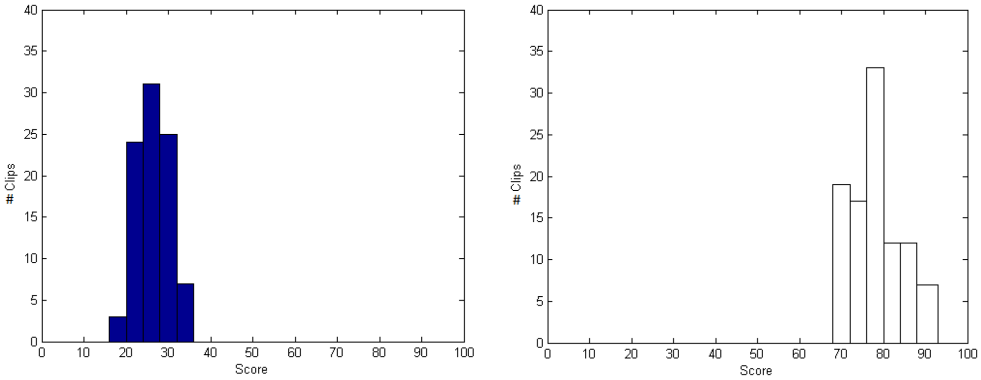
We obtained accuracies of 85.2% and 97.8% for DB-1 and DB-2-A, respectively. Although there is a greater number of errors occurring when the system deals with the accompanied singing recordings, its impact on the pitch-based rating is not fatal in the following experiment.

We then used the singing recordings in DB-2A to evaluate the performance of the proposed pitch-based rating method. Here, we set the parameters ε in Eq. (8) to be 0.5. In Eq. (10), the parameters k_1 and k_2 were determined to be 1.07 and –0.06, respectively, using a regression analysis on the human ratings for DB-2B. The results of human rating and system rating are listed in Table 4. Here, each singer's score was obtained by averaging the scores of his/her 15 recordings and then rounding off to an integer. We further ranked all the singers' scores in descending order. It can be seen from Table 4 that the ranking results obtained with our system are roughly consistent with those of the human rating, though there are score differences between the system rating and human rating. The results indicate that the singers in different groups can be well distinguished by our system.

Table 4. Results of the pitch-based rating for the 10 singers in DB-2A.

Singer Index		1	2	3	4	5	6	7	8	9	10
Group		I	I	II	II	II	II	II	II	III	III
Human Rating	Score	93	90	87	70	82	80	79	75	66	69
	Ranking	1	2	3	8	4	5	6	7	10	9
System Rating	Score	88	85	83	74	81	77	73	78	67	66
	Ranking	1	2	3	7	4	6	8	5	9	10

We further simulated the case that a singer performs a song irrelevant to the reference song clip by computing the distances between each pair of distinct song clips' note sequences and then substituting the distances into Eq. (10) to obtain the scores. Fig. 6 (a) shows the distribution of the resulting scores. We can see from Fig. 6 that the resulting scores are quite low if singers perform wrong songs, compared to the case in Fig. 6 (a) that singers perform correct songs. This result also implies that when the score of a test singing sample is less than 40, the singing may sound as if a wrong song is performed.



(a) Singers perform wrong songs. (b) Singers perform correct songs.

Fig. 6. Distribution of the pitch-based scores when singers perform correct and wrong songs.

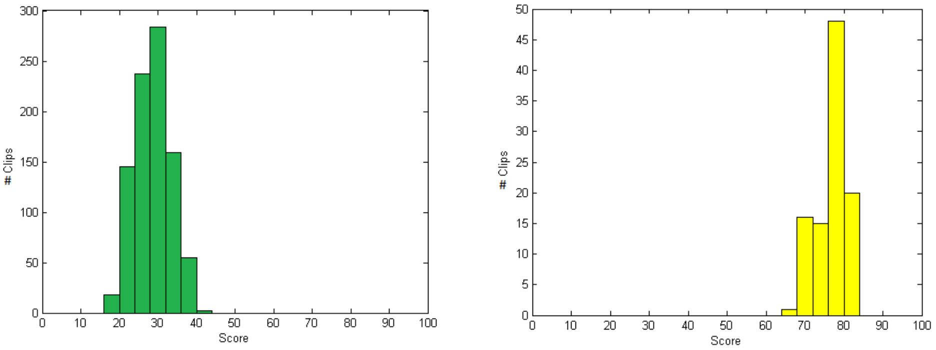
3.2.2 Experiments on volume-based rating

The validity of the volume-based rating were then examined. The parameters q_1 and q_2 in Eq. (12) were determined to be 1.02 and –0.17, respectively, using a regression analysis on the human ratings for DB-2B. The results of human rating and system rating are listed Table 5. We can see from Table 5 that the ranking results obtained with our system are roughly similar to those of the human rating.

Again, we simulated the case that a singer performs a wrong song clip. For each song clip in DB-1, the system used its energy sequence as a reference basis and then rated the 14 singing recordings in DB-2A that are irrelevant to the song of the reference basis. Fig. 7 shows the distribution of the resulting scores. It is clear from Fig. 7 that the resulting scores are quite low if singers perform wrong songs, compared to the case in Fig. 7 (a) that singers perform correct songs. Such a low score indicates that the proposed volume-based rating can well recognize if a singer performs a wrong song.

Table 5. Results of the volume-based rating for the 10 singers in DB-2A.

Singer Index		1	2	3	4	5	6	7	8	9	10
Group		I	I	II	II	II	II	II	II	III	III
Human Rating	Score	81	90	83	74	76	79	87	84	65	68
	Ranking	5	1	4	8	7	6	2	3	10	9
System Rating	Score	85	88	76	73	71	79	81	74	63	64
	Ranking	2	1	5	7	8	4	3	5	10	9



(a) Singers perform wrong songs.

(b) Singers perform correct songs.

Fig. 7. Distribution of the volume-based scores when singers perform correct and wrong songs.

3.2.3 Experiments on rhythm-based rating

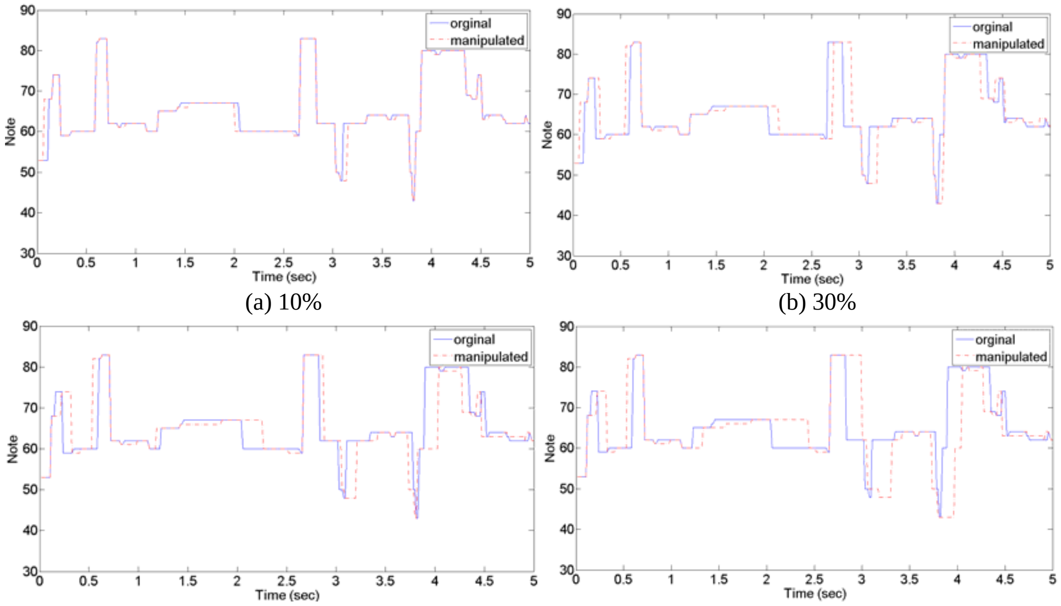
Next, we examined the validity of the rhythm-based rating. The parameters r_1 and r_2 in Eq. (13) were determined to be 1.04 and –0.08, respectively, using a regression analysis on the human ratings for DB-2B. Table 6 shows the rating results. We can see from Table 6 that the ranking results obtained with our system are roughly consistent with the human rating. To gain insight into the discriminability of our system to different levels of errors in rhythm, we randomly chose a singing clip performed by Singer #1 and manipulated its note sequence to simulate and measure how the score could drop when various levels of errors occur in rhythm, where Singer #1 is considered to be the one with the best singing capability among others in our database. Suppose that the original note sequence is (62,62,62,62,71,71,71,71,71,65,65,65). The manipulation is done by introducing two types of errors in the sequence, one is "ahead of a beat" like (62,62,71,71,71,71,71,71,71,65,65,65), and the other is "behind a beat" like (62,62,62,62,62,62,71,71,71,65,65,65). Fig. 8 shows some examples of introducing the rhythmic errors in a note sequence, where the percentages are calculated by

$$\frac{\text{Number of notes inserted/substituted/deleted in the original sequence}}{\text{Number of notes in the original sequence}} \times 100\%.$$

Fig. 9 shows the resulting drop in the score when rhythmic errors are introduced in a note sequence artificially. We can see from Fig. 9 that 10% errors roughly result in a drop of score by 3, and 50% errors can lead to a drop of score by 50. The results indicate that the proposed rhythm-based rating is capable of detecting the tiny rhythmic differences. This confirms the validity of the proposed rhythm-based rating.

Table 6. Results of the rhythm-based rating for the 10 singers in DB-2A.

Singer Index		1	2	3	4	5	6	7	8	9	10
Group		I	I	II	II	II	II	II	II	III	III
Human Rating	Score	90	87	83	80	87	72	77	81	70	79
	Ranking	1	2	4	6	3	9	8	5	10	7
System Rating	Score	89	88	82	85	84	77	79	74	71	73
	Ranking	1	2	5	3	4	7	6	8	10	9



(a) 10% (b) 30%

(c) 50% (d) 70%

Fig. 8. Examples of introducing the rhythmic errors in a note sequence.

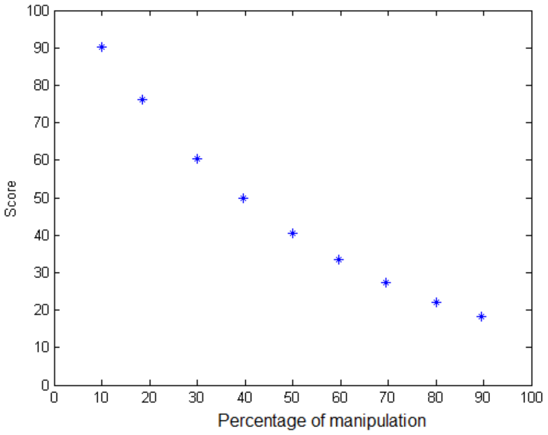


Fig. 9. Drop in score when rhythmic errors are introduced in a note sequence artificially.

3.2.4 Combination of pitch-based, volume-based, and rhythm-based rating

Lastly, the overall rating system using Eq. (1) was evaluated. Here, the weights w_1, w_2, and w_3 were estimated to be 0.44, 0.16, and 0.40, respectively, using the least square analysis of the human ratings for DB-2B. Table 6 lists the overall rating results. We can see from Table 7 that the scores obtained with the system rating roughly match those of the human rating. To evaluate the consistency between the results of the system rating and human rating, we computed the Pearson product-moment correlation coefficients [15] between human rating and system rating. As shown in Tables 8, we can see that there is a high positive correlation between the human rating and our system rating. The results obtained with our system are also comparable to those in a previous work [13] using Karaoke music as reference bases, where the latter's task is considered to be easier than that of this work. This indicates that our system is capable of exploiting pitch, volume, and rhythm-based features from CD/mp3 song recording as reference bases for singing performance evaluation.

Table 7. Overall rating based on Eq. (1).

Singer Index		1	2	3	4	5	6	7	8	9	10
Group		I	I	II	II	II	II	II	II	III	III
Human Rating	Score	90	89	85	75	83	77	80	79	67	73
	Ranking	1	2	3	8	4	7	5	6	10	9
System Rating	Score	88	87	81	78	81	77	77	76	68	68
	Ranking	1	2	3	5	4	7	8	6	10	9

Table 8. The Pearson product-moment correlation coefficient between the human rating and system rating.

Rating Method	Our System	System in [13]
Pitch-based Rating	0.79	0.80
Volume-based Rating	0.76	0.77
Rhythm-based Rating	0.84	0.86
Overall Rating	0.80	0.82

4. CONCLUSIONS

This study has developed an automatic singing evaluation system for general public. Given a CD/mp3 song recording as the reference basis, the proposed system rates a user's singing performance by comparing it with the vocal in the song recording. This modality allows users to not only enjoy listening to and singing with CD/mp3 songs but also know how well or bad they sing. Recognizing a majority of songs contain background accompaniments during most or all vocal passages, we propose methods to extract pitch-, volume-, and rhythm-based features of the original singer in the accompanied vocals by reducing the interferences from background accompaniments. After examining the consistency between the results of automatic singing evaluation with the subjective judgments of musicians, we showed that the proposed system is capable of providing singers with a reliable rating.

In the future, we will consider timbre-based analysis and sung lyrics verification to further improve the singing evaluation system. In the timbre-based analysis, we may consider to use vibrato as a cue of singing evaluation. The method developed in [1] could be incorporated into our system. With regard to sung lyrics verification, there would be a need to investigate the difference between speech and singing so that a speech recognition system can be adapted to handle singing performances. In addition, rhythm-based rating may be further improved by incorporating note onset detection into our system. However, it is a prerequisite to develop reliable algorithms for detecting the onsets of notes sung in the accompanied vocal recordings.

REFERENCES

1. T. Nakano, M. Goto, and Y. Hiraga, “An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features,” in *Proceedings of International Conference on Spoken Language Processing*, 2006, pp. 1706-1709.
2. T. Nakano, M. Goto, and Y. Hiraga, “Subjective evaluation of common singing skills using the rank ordering method,” in *Proceedings of International Conference on Music Perception and Cognition*, 2006, pp. 1507-1512.
3. T. Nakano, M. Goto, and Y. Hiraga, “MiruSinger: a singing skill visualization interface using real-time feedback and music CD recordings as referential data,” in *Proceedings of IEEE International Symposium on Multimedia*, 2007, pp. 75-76.
4. P. Lal, “A comparison of singing evaluation algorithms,” in *Proceedings of International Conference on Spoken Language Processing*, 2006, pp. 2298-2301.
5. O. Mayor, J. Bonada, and A. Loscos, “Performance analysis and scoring of the singing voice,” in *Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing*, 2009.
6. J. G. Hong and U. J. Kim, “Performance evaluator for use in a karaoke apparatus,” US Patent No. 5,557,056, 1996.
7. C. S. Park, “Karaoke system capable of scoring singing of a singer on accompaniment thereof,” US Patent No. 5,567,162, 1996.
8. K. S. Park, “Performance evaluation method for use in a karaoke apparatus,” US Patent No. 5,715,179, 1998.
9. B. Pawate, “Method and system for karaoke scoring,” US Patent No. 5,719,344, 1998.
10. T. Tanaka, “Karaoke scoring apparatus analyzing singing voice relative to melody data,” United States Patent, 5,889,224, 1999.
11. H. M. Wang, “Scoring device and method for a karaoke system,” US Patent No. 6,326,536, 2001.
12. P. C. Chang, “Method and apparatus for karaoke scoring,” US Patent No. 7,304,229, 2007.
13. W. H. Tsai and H. C. Lee, “Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features,” *IEEE Transactions on Audio, Speech, Language Processing*, Vol. 20, 2012, pp. 1233-1243.
14. H. M. Yu, W. H. Tsai, and H. M. Wang, “A query-by-singing system for retrieving karaoke music,” *IEEE Transactions on Multimedia*, Vol. 10, 2008, pp. 1626-1637.

15. R. A. Fisher, "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron*, Vol. 1, 1921, pp. 3-32.
16. J. P. Bello, L. Daudet, S. Abdullah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech Audio Processing*, Vol. 13, 2005, pp. 1035-1047.
17. S. Bock and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proceedings of the 16th International Conference on Digital Audio Effects*, 2013, pp. 55-61.

Wei-Ho Tsai (蔡偉和) received his B.S. degree in Electrical Engineering from National Sun Yat-Sen University, Kaohsiung, Taiwan, in 1995. He received his M.S. and Ph.D. degrees in Communication Engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1997 and 2001, respectively. From 2001 to 2003, he was with Philips Research East Asia, Taipei, Taiwan, where he worked on speech processing problems in embedded systems. From 2003 to 2005, he served as a Postdoctoral Fellow at the Institute of Information Science, Academia Sinica, Taipei, Taiwan. He is currently a Professor in the Department of Electronic Engineering and Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taiwan. His research interests include spoken language processing and music information retrieval. Dr. Tsai is a life member of ACLCLP and a member of IEEE.

Cin-Hao Ma (馬勤皓) received the B.S. degree in Electronic Engineering from National Taipei University of Technology, Taipei, Taiwan, in 2012. He is pursuing the Ph.D. degree in Computer and Communication Engineering at National Taipei University of Technology currently. His research interests include signal processing and multimedia applications.

Yi-Po Hsu (徐毅博) received his M.S. degree in Computer and Communication Engineering from National Taipei University of Technology, Taipei, Taiwan, in 2014. His research interests include signal processing and multimedia applications.