Automatic Evaluation of Karaoke Singing Based on Pitch, Volume, and Rhythm Features

Wei-Ho Tsai, *Member, IEEE*, and Hsin-Chieh Lee

***Abstract*—This study aims to develop an automatic singing evaluation system for Karaoke performances. Many Karaoke systems in the market today come with a scoring function. The addition of the feature enhances the entertainment appeal of the system due to the competitive nature of humans. The automatic Karaoke scoring mechanism to date, however, is still rudimentary, often giving inconsistent results with scoring by human raters. A cause of blunder arises from the fact that often only the singing volume is used as the evaluation criteria. To improve on the singing evaluation capabilities on Karaoke machines, this study exploits various acoustic features, including pitch, volume, and rhythm to assess a singing performance. We invited a number of singers having different levels of singing capabilities to record for Karaoke solo vocal samples. The performances were rated independently by four musicians, and then used in conjunction with additional Karaoke Video Compact Disk music for the training of our proposed system. 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 0.82.**

***Index Terms*—Accompaniment, Karaoke, singing evaluation, solo vocal.**

I. Introduction

KARAOKE is a popular recreational pastime in East Asia. With a microphone Karaoke machine, people sing along with onscreen guidance to recordings of popular songs from which the vocals have been removed. In addition to serving as a form of entertainment for amateur singers, Karaoke is a convenient way to help people practice singing. A Karaoke machine with an intelligent singing evaluation capability, therefore, is a useful tool to provide singers with an immediate feedback.

Although many Karaoke apparatuses or games [1], [2] come with an automatic scoring feature, their evaluation capabilities, however, do not always match that of human evaluation. A cause of blunder arises from the fact that an automatic singing evaluation method has not been thoroughly investigated. A vast majority of Karaoke apparatuses today use loudness as the only criteria for performance evaluation. Some apparatuses even generate a score randomly for fun, which is highly misleading to the singers. To provide karaoke singers with useful feedbacks on their performances, a more robust scoring method is needed.

This study aims to explore various acoustic features, including pitch, volume, and rhythm, to assess a singing performance. We invited a number of singers having various levels of singing capabilities for recordings of solo Karaoke performances. The performance samples were rated by four musicians, and then used in conjunction with Karaoke Video Compact Disk (VCD) music for the training of our proposed system. Our experiment shows that the results of automatic singing evaluation are close to the human rating.

The remainder of this paper is organized as follows. In Section II, we formulate the problem of singing performance evaluation and discuss the reference basis for evaluation. Section III reviews the related work and techniques. Section IV introduces our system design approach. Section V discusses our experiment results. Then, in Section VI, we present our conclusions and indicate the directions of our future work.

II. Problem Overview

Given a Karaoke singing performance, the aim of our system is to assess the performer's singing ability in terms of technical accuracy and assign a rating score. Depending on the subject matter expertise and the judging criteria, evaluating singing skills can be highly subjective. As a preliminary basis for the goal of developing an automatic singing skill evaluation, this study assesses a performer's singing ability in terms of technical accuracy in pitch, rhythm, and dynamics. The scope of this application is suitable for use as an entertainment and learning tool.

References for judging pitch, rhythm, and dynamics accuracy are required for the automatic scoring system. Depending on the types of Karaoke apparatuses, a reference basis can be derived from the following sources:

- **Music scores and lyrics**. If the music score and lyrics information for every song are available, the evaluation task can be accomplished by checking the precision of a singer's pitch, timing, rhythm, and articulation against the score and lyrics. However, most Karaoke apparatuses today do not have access to music scores; therefore, it is difficult to build a singing evaluation system based on this approach.
- **MIDI (Musical Instrument Digital Interface) files**. If there is a "solo track" present in a MIDI file, we can then examine a Karaoke performance against the musical information contained in the solo track. However, most MIDI-based Karaoke systems today do not have access to a "solo

Manuscript received November 30, 2010; revised July 29, 2011; accepted October 10, 2011. Date of publication November 18, 2011; date of current version February 17, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sylvain Marchand.

The authors are with the Department of Electronic Engineering and Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei City 10608, Taiwan (e-mail: whtsai@ntut.edu.tw; t9419004@ntut.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2174224

1558-7916/$31.00 © 2011 IEEE

track" or any lyrics information. The MIDI files approach, therefore, is not presently feasible.

- **CD music**. To compare the Karaoke singing with the vocal from the original music recording, the evaluation system would need to extract the vocals from polyphonic music. However, as extracting vocals from polyphonic music is known to be a very difficult problem, building such a singing evaluation system would be a big challenge.
- **Karaoke VCD music**. Unlike a regular music CD, where the stereo recording stores two similar audio channels, a Karaoke VCD encompasses two distinct channels. One is a mixture of the lead vocals and background accompaniment, and the other consists of the accompaniment only. Although the lead vocal is not recorded on a separate track without the accompaniment, the recording format makes it easier to extract and exploit the vocals in a Karaoke VCD than in a regular music CD.
- **Solo vocal track**. In certain Karaoke systems such as the digital video systems (DVS) or the laser disc (LD) karaoke system in Japan, the artist's voice is recorded on a dedicated vocal track. Given the vocal data, a direct comparison can be made between the voice of the original artist and the Karaoke singing. However, as most Karaoke apparatuses do not have a separate track dedicated to vocals only, using this approach would require tremendous effort to collect the vocal data.

Among the five reference sources mentioned above, music CD is the most popular storage format for music performance data. Thus, a singing evaluation system based on such format would be most useful. However, due to the difficulty in separating vocals from the accompaniment, this study focuses on developing a system based on music data stored in Karaoke VCD.

In addition, a singing scoring system must exploit various acoustic cues to make a comparison between a test singing piece and the reference basis. From the standpoint of the elements of music, the basic acoustic cues are:

- **Volume**, which reflects the intensity of sound in a piece of music;
- **Pitch**, which refers to the actual value of the note sung;
- **Rhythm**, which relates to the timing of the musical sound and silences;
- **Timbre**, which describes the quality (e.g., from dull to lush) or color (e.g., from dark to bright) of a tone produced by a singer.

In general, volume, pitch, and rhythm are much related to whether or not a song is performed correctly, whereas timbre mainly involves natural voice characteristics of individuals and therefore is hard to be exploited as a fair cue to assess a singing performance. However, since songs performed with the same tune but different lyrics mainly differ in their timbres, it might be necessary to perform timbre-based analysis, when trying to examine if the lyrics performed by a singer are correct. Nevertheless, this study does not intend to deal with the examination of sung lyrics, because this problem is too difficult to handle well at current stage. Moreover, as most Karaoke machines have onscreen lyrics, singers would easily know if their sung lyrics are incorrect, without relying on the automatic evaluation system.

Another cue that can be derived from pitch is the presence of *vibrato* in singing. Vibrato is a slight variation of pitch resulting from the oscillation of the vocal cords. Some singers use vibrato to enhance the expressiveness of their performance. For example, vibrato is commonly used to place emphasis on significant words or phrases of a musical piece. Singing vibrato, however, is an acquired vocal technique and usually requires years to master. Some singers have an overly fast vibrato, called *tremolo*, while others have a wide and slow vibrato, called *wobble*. Since neither the tremolo nor the wobble is the desired vocal technique, vibrato, thus, can be considered as an important cue to distinguish between a well-trained singer and a mediocre singer [4].

Although singing vibrato is often considered an artistically expressive singing attribute, vibrato in tones can become problematic in choral singing, however. Choral directors sometimes ask the chorus to sing with a straight tone. This is because vibrato varies from singer to singer, which makes it difficult for a chorus to produce a harmonious blend of sounds. Since vibrato is not notated in musical scores, which makes it a subjective measure of assessment, this work does not use it to evaluate a singing performance. Nevertheless, it should be noted that vibrato would be a crucial feature in designing an automatic system capable of discriminating the superiority of a singer over another.

In addition to the acoustic cues discussed above, prior work [15]–[17] showed other examples of quantitative measurements useful to distinguish between classically trained and untrained singers. One of the most prominent measurements is the singing power ratio (SPR) [15]. SPR is the ratio of the highest spectral peak between 2 and 4 kHz and the highest spectral peak between 0 and 2 kHz in sustained vowels or vocalic segments. Lower SPR indicates greater energy in the higher harmonics, which results in the "ringing" voice quality most perceptible in *bel canto*, a virtuosic, operatic style of singing. In general, most untrained singers will have higher SPR than classically trained singers because a bel canto style of "ringing" quality in singing is not easily achievable. Higher SPR in a karaoke performance, however, does not necessarily mean the performance is less impressive than those with lower SPR, because the operatic style of singing, one may argue, is unfit and undesirable for a karaoke style of singing. This study, thus, does not consider SPR as a parameter in designing our singing evaluation system, but focuses on using the acoustic cues that are notated in musical scores.

III. Related Work

Up to now, most of the singing-evaluation studies [3]–[14] are reported in patent documentation. Only very few studies are reported in scientific literature. Table I summarizes a list of related studies. Most of these patents describe their implementation details; however, they do not discuss the rationale of their evaluation method. In addition to their lack of theoretical foundation, most of these patents failed to show results of their experiments or any qualitative analysis conducted to validate their methods.

In the scientific literature, Nakano *et al.* [3] explore the criteria that human subjects use in the singing evaluation. The issue of how to design an automatic singing-evaluation system is left

TABLE I
RELATED STUDIES ON AUTOMATIC SINGING PERFORMANCE EVALUATION

Research Unit	Publication	Year	Reference Basis	Acoustic Cues
Daewoo Electronics [8]	US Patent No. 5,557,056	1996	"Karaoke Music". However, it is not clear whether the music is accompaniment only or accompanied singing.	Volume
Daewoo Electronics [9]	US Patent No. 5,567,162	1996	Pure Solo Singing	Spectrum Differences
Daewoo Electronics [10]	US Patent No. 5,715,179	1998	"Karaoke Music". However, it is not clear whether the music is accompaniment only or accompanied singing.	Waveform Differences
Texas Instrument [11]	US Patent No. 5,719,344	1998	Pure Solo Singing	Volume
Yamaha [12]	US Patent No. 5,889,224	1999	MIDI	Volume & Pitch
Winbond [13]	US Patent No. 6,326,536	2001	Lowpass-filtered CD Music	Volume
University of Tsukuba and *AIST* [4]	Interspeech	2006	Pure Solo Singing	Pitch & Vibrato
University of Edinburgh [6]	Interspeech	2006	Pure Solo Singing	Pitch
University of Tsukuba and *AIST* [5]	IEEE Symp. Multimedia	2007	CD Music	Pitch & Vibrato
Mediatek [14]	US Patent No. 7,304,229	2007	"Reference Vocal Input". However, it is not clear whether the reference vocal is pure solo singing or accompanied singing.	Pitch
University Pompeu Fabra and *BMAT* [7]	AES 35th Int. Conf.	2009	MIDI	Pitch

to another work of theirs [4]. In [4], a 2-class (good/poor) classifier based on support vector machine is built to determine which class a test singing sample belongs to. The acoustic features used in the classifier are pitch interval accuracy and vibrato. Later on, Nakano *et al.* [5] develop a singing skill visualization interface, *MiruSinger*, which analyzes and visualizes the pitch contours of a test singing sample and the vocal-part in music CD recordings. On the other hand, Lal [6] proposes two pitch-based similarity measures to determine how close a user's singing clip is to the reference singing clip with no background music. In [7], Mayor *et al.* propose a method to rate the performance of a singer by aligning it to a reference MIDI. Although the above-mentioned works have provided better solutions than existing commercial singing-evaluation systems, most of them only consider pitch-based cues and present a preliminary experiment results.

Another research topic closely related to singing evaluation is the computer assisted singing training [18], [19], which resembles the pedagogic research on pronunciation training [20] and second-language learning [21], [22]. Several tools have been developed to help learners improve their singing performance based on real-time visual feedback. The tools guide a user to sing the prompted notes and then display the acoustic information, such as waveform, pitch, spectrum, vowel identity, and vocal tract area, with respect to the user's singing. However, the purpose of developing these tools lies more on assisting teachers than replacing teachers; hence, learners still rely on teachers' comments to know what the wrong and right with their singings are. The problem of how to evaluate a singing performance automatically is not investigated in these studies.

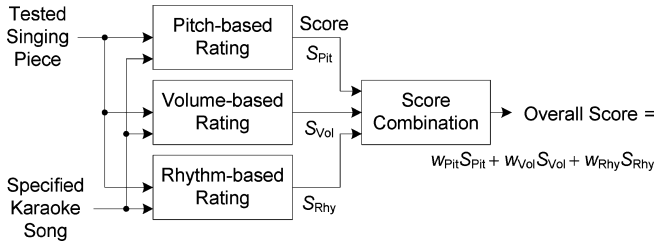

Fig. 1. Proposed singing-evaluation system.

IV. METHODOLOGY

Fig. 1 shows the proposed singing-evaluation system. When a singing piece is evaluated, the system performs volume-based rating, pitch-based rating, and rhythm-based rating, using the Karaoke song which the singer sings as a reference basis. The resulting scores from each component are then combined using a weighted sum method:

$$\text{Overall Score} = w_{\text{Pit}} \cdot S_{\text{Pit}} + w_{\text{Vol}} \cdot S_{\text{Vol}} + w_{\text{Rhy}} \cdot S_{\text{Rhy}} \quad (1)$$

where S_{Pit}, S_{Vol}, and S_{Rhy} are the scores obtained with pitch-based rating, volume-based rating, and rhythm-based rating, respectively; w_{Pit}, w_{Vol}, and w_{Rhy} are the adjustable weights that sum to 1.

A. Pitch-Based Rating

Pitch refers to the relative lowness or highness that we hear in a sound. To sing in tune, a sequence of notes must be sung in the correct pitch along with the appropriate duration. This study uses the MIDI note scale to compare the sequence of notes sung in an evaluated recording with the ones sung in the reference recording.

The rating begins by converting the waveform of a singing recording into a sequence of MIDI notes. Let n_m, $1 \le m \le M$, be the inventory of possible notes performed by a singer. Thus, our aim is to determine which among the M possible notes is most likely sung at each instant. We apply the strategy in [23] to solve this problem. First, the vocal signal is divided into frames by using a P-length sliding Hamming window, with $0.5P$-length overlapping between frames. Every frame then undergoes a fast Fourier transform (FFT) with size J. Let $x_{t,j}$ denote the signal's energy with respect to FFT index j in frame t, where $1 \le j \le J$, and $x_{t,j}$ has been normalized to the range between 0 and 1. Then, the signal's energy on mth note in frame t can be estimated by

$$\hat{x}_{t,m} = \max_{\forall\, j, U(j)=n_m} x_{t,j} \quad (2)$$

and

$$U(j) = \left\lfloor 12 \cdot \log_2 \left(\frac{F(j)}{440} \right) + 69.5 \right\rfloor \quad (3)$$

where $\lfloor\ \rfloor$ is a floor operator, $F(j)$ is the corresponding frequency of FFT index j, and $U(\cdot)$ represents a conversion between the FFT indices and the MIDI note numbers.

Ideally, if note n_m is sung in frame t, the resulting energy, $\hat{x}_{t,m}$, should be the maximum among $\hat{x}_{t,1}, \hat{x}_{t,2}, \ldots, \hat{x}_{t,M}$. However, it is sometimes the case that the energy of a sung note is smaller than that of its harmonic note. To avoid the interference of harmonics in the estimation of sung notes, we use the strategy of sub-harmonic summation (SHS) [24], which computes a value for the "strength" of each possible note by summing the signal's energy on a note and its harmonic note numbers. Specifically, the strength of note n_m in frame t is computed using

$$y_{t,m} = \sum_{c=0}^{C} h^c \hat{x}_{t,m+12c} \tag{4}$$

where C is the number of harmonics considered, and h is a positive value less than 1 that discounts the contribution of higher harmonics. The result of summation is that the sung note usually receives the largest amount of energy from its harmonic notes. Thus, the sung note in frame t can be determined by choosing the note number associated with the largest value of the strength. However, recognizing that a sung note usually lasts several frames, the decision could be made by including the information from neighbor frames. Specifically, we determine the sung note in frame t by choosing the note number associated with the largest value of the strength accumulated for adjacent frames, i.e.,

$$o_t = \arg \max_{1 \le m \le M} \sum_{b=-B}^{B} y_{t+b,m}. \tag{5}$$

Further, the resulting note sequence is refined by taking into account the continuity between frames. This is done with median filtering, which replaces each note with the local median of notes of its neighboring $\pm B$ frames, to remove jitters between adjacent frames. In the implementation, the range of n_m is set to be $30 \le n_m \le 90$. However, considering the normal range of notes in popular songs, only the notes between 43 and 83 are regarded as the possible sung notes. For the notes falling outside this range, they are regarded as consonants or pauses and replaced by a fixed value of 40.

In addition, the above method, however, 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 Karaoke 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 apply Spectral Subtraction (SS) to reduce the background interference. As mentioned earlier, Karaoke music encompasses two distinct channels in each track: one is a mixture of the lead vocals and background accompaniment, and the other consists of accompaniment only. Although the two audio channels are distinct, the music in the accompaniment-only channel usually sounds similar to the background accompaniment in the accompanied vocal channel. By subtracting accompaniment-only channel's spectrum from accompanied vocal channel's spectrum, an approximated solo singing spectrum could be obtained. However, as the volume in the accompanied vocal channel may not be always larger than that of the accompaniment-only channel, direct subtraction could result in a negative-value spectrum. To overcome this problem, we use a weighted subtraction strategy stemming from [25]. Fig. 2 shows an example of Karaoke music after performing SS. From Fig. 2(c), we can see that the fundamental frequencies of the singing become rather visible, as SS reduces the accompaniment in the accompanied vocal channel substantially.

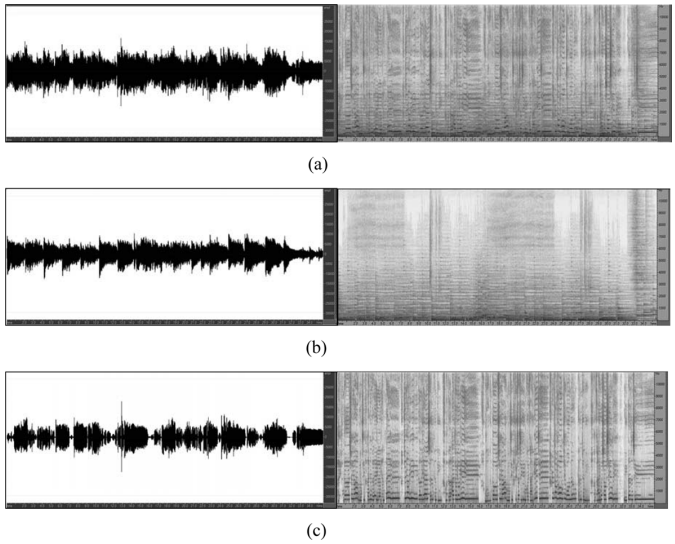

Fig. 2. Example of Karaoke music after performing spectral subtraction (SS). (a) Waveform and spectrogram of the accompanied vocal channel. (b) Waveform and spectrogram of the accompaniment-only channel. (c) Waveform and spectrogram of the SS result.

Fig. 3 shows the block diagram of the pitch-based rating. In the offline phase, a Karaoke song's accompanied vocal signal is converted from its waveform representation $v[n]$ into a reference note sequence $\mathbf{O} = \{o_1, o_2, \ldots, o_T\}$. Since $v[n]$ contains background accompaniment $a[n]$, which approximates the accompaniment-only channel's signal $a'[n]$, SS is performed prior to the note sequence generation. In the online phase, a singing recording is converted from its waveform signal $s'[n]$ into a note sequence $\mathbf{O'} = \{o'_1, o'_2, \ldots, o'_{T'}\}$. Then, the performer's singing skill is evaluated on the basis of the distance between $\mathbf{O}$ and $\mathbf{O'}$. 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 $\mathbf{O}$ and $\mathbf{O'}$.

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

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

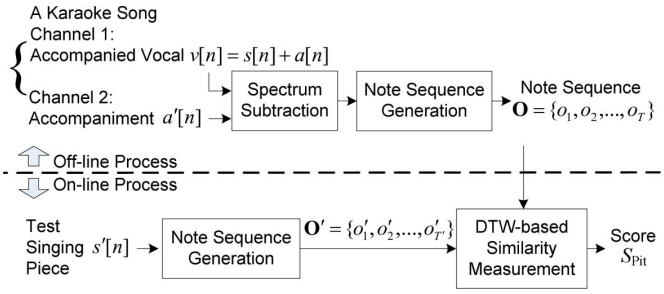

Fig. 3. Pitch-based rating.

and

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

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

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

After the distance matrix $\mathbf{D}$ is constructed, the distance between $\mathbf{O}$ and $\mathbf{O}'$ can be evaluated by

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

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 $\text{Dist}(\mathbf{O},\mathbf{O}')$ is then converted to a score between 0 and 100:

$$S_{\text{Pit}} = 100 \cdot k_1 \exp[k_2 \cdot \text{Dist}(\mathbf{O},\mathbf{O}')] \tag{10}$$

where k_1 and k_2 are tunable parameters used to control the distribution of S_{Pit}.

B. Volume-Based Rating

When a song is composed, abbreviations or symbols called dynamics are notated in music scores to indicate the degree of loudness or softness of a piece of music, and whether there is a change in volume. Dynamics are relative, rather than absolute. They only indicate that music in a passage so marked should be a little louder or a little quieter. Thus, interpretations of dynamic levels are left mostly to the performer. Despite this, there should be a similar pattern of volume variations across time, when different singers perform the same song.

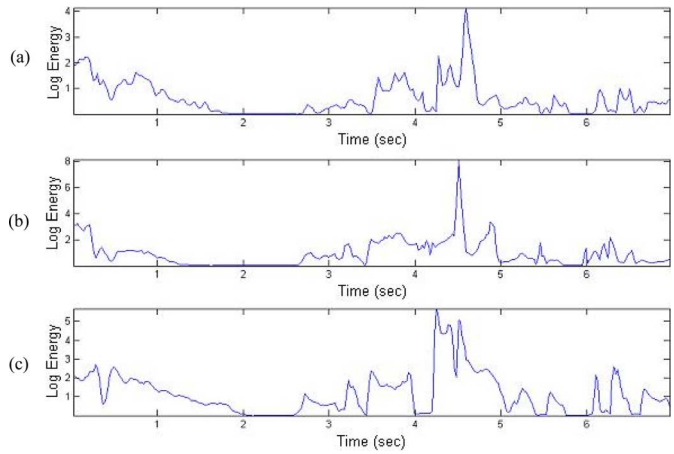

Fig. 4. Example of a Mandarin song performed by three part-time pub singers, where each figure represents the normalized short-term log-energy contour of a singer's performance.

Fig. 4 shows an example of a Mandarin song performed by three part-time pub singers. Here, waveforms are recorded in a quiet room without any background music[1], and are then converted into a sequence of short-term log-energy using 30-ms sliding windows with 15-ms advance length. The sequences are further normalized by removing the individual means. We can see that the contours of the normalized log-energy sequences in Fig. 4(a)–(c) are similar. This serves as a basis for the proposed volume-based rating.

Fig. 5 shows the processes of the volume-based rating. As Karaoke VCD music does not contain solo singing, direct comparison of energy sequence between a test singing and its reference singing is infeasible. To solve this problem, we estimate the short-term log-energy sequence of the reference singing using the signal resulting from the spectrum subtraction. In addition, to exclude the tempo variations that may affect the volume-based rating, we apply DTW to measure the distance, $\text{Dist}(\mathbf{E},\mathbf{E}')$, between sequence of the reference singing, $\mathbf{E}$, and sequence of the test singing, $\mathbf{E}'$. Then, a volume-based score is obtained using

$$S_{\text{Vol}} = 100 \cdot q_1 \exp[q_2 \cdot \text{Dist}(\mathbf{E},\mathbf{E}')] \tag{11}$$

where q_1 and q_2 are tunable parameters used to control the distribution of S_{Vol}.

C. Rhythm-Based Rating

Rhythm is related to the timing of musical sound and silences performed by a singer. Although every song has a standard rhythm, performers sometimes take the liberty of the time to elicit certain emotional responses in the listeners. In Karaoke, since the accompaniment is prerecorded, the singer must follow the pace of the accompaniment. If they do not follow the flow of the accompaniment, then the performance may sound out of beat. Thus, our basic idea of rhythm-based rating is to evaluate

[1]See Section V-A for more the details of music data.

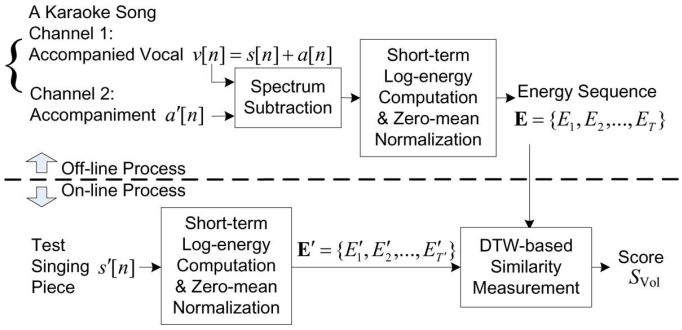

Fig. 5. Volume-based rating.

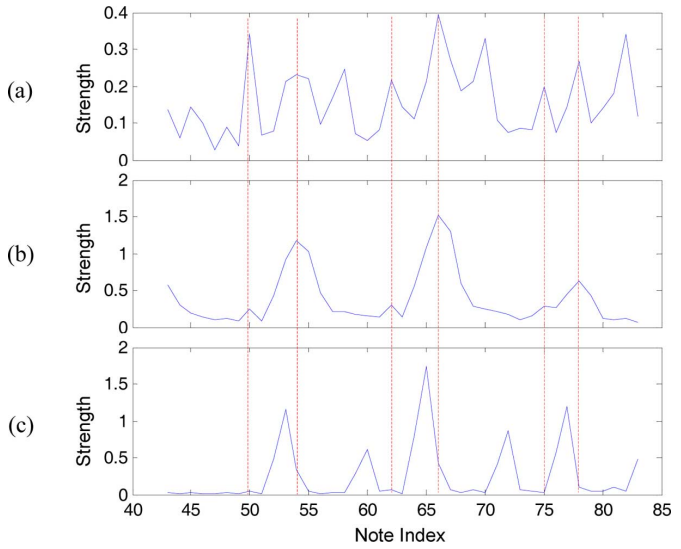

Fig. 6. (a) note strengths of an accompaniment signal in a certain frame, (b) note strengths of a singing signal synchronous with the accompaniment, and (c) note strengths of a singing signal asynchronous with the accompaniment.

for the synchronicity between the singing and the accompaniment, as singers often have a tendency to drag or rush at particular points of a song.

Fig. 6 shows an example of synchronous (in-beat) and asynchronous (out-of-beat) cases between singing and accompaniment, in which the accompaniment belongs to disco music and mainly contains bass, guitar, and electronic piano sounds produced by synthesizers. Fig. 6(a) represents the note strengths of an accompaniment signal in a certain frame, computed using (4). Fig. 6(b) represents the note strengths of a singing voice signal in the frame synchronous with the signal in (a). Fig. 6(c) represents the note strengths of a singing voice signal in the frame asynchronous with the signal in (a). We can see that Fig. 6(a) and (b) has many peaks (indicated by the dotted lines) in the same note indices. Such consistency is probably the reason for why a singing voice with correct rhythm to the accompaniment sounds harmonious. On the contrary, we can see that Fig. 6(a) and (c) has no consistent peaks in the same note indices. The result of such inconsistency is that the two signals sound irrelevant to each other. Thus, it is reasonable to assume that "synchronous accompanied singing" can be distinguished from "asynchronous accompanied singing" by examining their note strength patterns.

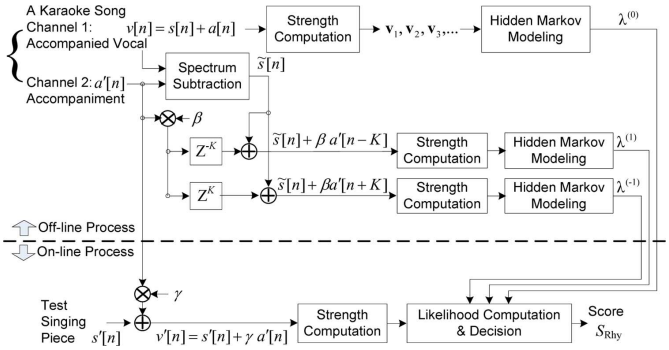

Fig. 7. Rhythm-based rating.

Fig. 7 shows the block diagram of the proposed rhythm-based rating. Our basic strategy is to represent synchronous and asynchronous accompanied singing by probabilistic models and then perform stochastic recognition. For each song, an "in-beat model" is built using the accompanied vocal signal extracted from Karaoke VCD music. The signal is first converted from its waveform into a sequence of strength vectors using (4), and then represented by a hidden Markov model (HMM). Specifically, the observations for the HMM are a sequence of vectors $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_t, \ldots\}$, where $\mathbf{Y}_t = [y_{t,1}, y_{t,2}, \ldots, y_{t,M}]^{\text{Tr}}$, and $y_{t,m}$, $1 \leq m \leq M$, is the strength of note n_m in frame t. The distribution of each observation in each state of the HMM is a mixture of Gaussian densities. Parameters of HMM, including initial probabilities, state transition probabilities, mixture weights, mean vectors, and covariance matrices, are estimated using Baum–Welch algorithm [26]. We denote the in-beat HMM by $\lambda^{(0)}$.

As to asynchronous accompanied singing, we create two "out-of-beat HMMs" using manually mixed accompanied singing data. The first HMM models a performer singing ahead of a beat. It is trained in the following way. First, an approximated solo singing is extracted from the accompanied vocal channel using SS. The approximated solo singing is then superimposed with the accompaniment shifted to the right by K samples in the time domain. Thus, the resulting accompanied singing data sounds as if a performer always sings ahead of a beat. In order for the vocal-to-accompaniment ratio of a manually mixed accompanied singing be close to that of the true accompanied singing, the accompaniment is multiplied by a scale β before it is mixed with the approximated solo singing. The scale is determined in such a way that the energy of the manually mixed accompanied singing is equal to that of the accompanied vocal channel. Next, the data is converted into a sequence of strength vectors using (4). The sequence is then represented by an HMM using Baum–Welch algorithm. We denote this HMM by $\lambda^{(1)}$. On the other hand, the second HMM models a performer singing falling behind a beat. It is trained using the data generated by mixing the approximated solo singing and the accompaniment after being scaled by β and shifted to the left by K samples in the time domain. We denote this HMM by $\lambda^{(-1)}$.

Given a test singing recording, our system mixes it with the accompaniment scaled by a factor γ, according to an appropriate vocal-to-accompaniment ratio. The strength sequence

of the mixed sound is then computed and divided into several W-length non-overlapping segments $\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_L$. Next, the attribute of each segment is determined by

$$A_\ell = \arg \max_{-1 \le j \le 1} \Pr(\mathbf{G}_\ell | \lambda^{(j)}) \tag{12}$$

where $A_\ell = -1$, 0, and 1 represent that the singing is behind a beat, in beat, and ahead of a beat, respectively. As $(A_\ell = \pm 1)$ indicates the occurrence of incorrect rhythm, the system computes a rhythm-based score using

$$S_{\mathrm{Rhy}} = 100 \cdot \frac{1}{L} \sum_{\ell=1}^{L} \delta(A_\ell) \tag{13}$$

where $\delta(\cdot)$ is the Dirac delta function.

V. Experiments

A. Music Database

Our music data consists of two databases. The first database, denoted by DB-1, contains 20 Mandarin song clips extracted from Karaoke VCDs. Each clip ranges in duration from 25 to 40 seconds and contains a verse or chorus part of song. For computational efficiency, each extracted music track was downsampled from 44.1 kHz to 22.05 kHz and stored as PCM wave. The second database, denoted by DB-2, contains singing samples recorded by ourselves in a quiet room. We employed 25 singers to record for solo vocal parts of the 20 Mandarin song clips. The recordings were stored in mono PCM wave with 22.05-kHz sampling rate and 16-bit quantization level. When singers performed, the Karaoke accompaniments were output to a headset and were not captured in the recordings.

Among the 25 singers, 10 are considered to have good singing capabilities, in which most of them are part-time pub singers or have experiences in formal singing contests, e.g., One Million Star, in Taiwan. We marked the 10 singers by Group I. The other 10 among the 25 singers are those who like to sing Karaoke, but their singing capabilities are far from professional. We marked them by Group II. The remaining 5 among the 25 singers are considered to have poor singing capabilities. They sometimes cannot follow the tune, and some of them even never sing Karaoke before. We marked the 5 singers by Group III. In addition, to establish the ground truth for automatic singing evaluation, we employed four musicians to rate the singing recordings independently. The ratings were done in terms of technical accuracy in pitch, volume, rhythm, and combination thereof. We have also evaluated the consistency between the four musicians' ratings using the Pearson product-moment correlation coefficient [27]. The coefficient was first computed for each pair of musicians' ratings. Then, the resulting six coefficients were averaged. We obtained correlation coefficients of 0.83, 0.82, 0.87, and 0.86 between the four musicians' rating on pitch-based, volume-based, rhythm-based, and overall ratings, respectively. In addition, the rating results given by the four musicians were then averaged to form a reference score for each singing recording.

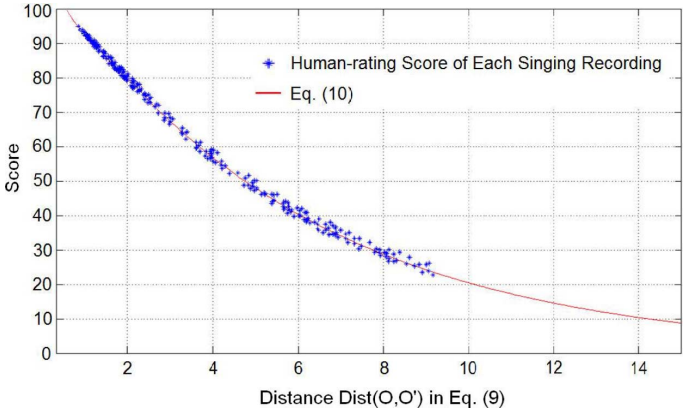

Fig. 8. Distribution of the human-rating scores based on pitch accuracy for the singing recordings in DB-2B, and the resulting regression curve.

Database DB-2 was further divided into two subsets. The first subset, denoted by DB-2A, was used to test our system. It 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, was used to tune the parameters in (1), (10) and (11). It contains the remaining recordings of DB-2 not covered in DB-2A.

B. Experiment Results

1) Experiments on Pitch-Based Rating: Before examining the validity of the pitch-based rating, our first experiment was conducted to investigate the distribution of score S_{Pit}. The length of frame and FFT size were set to be 30-ms and 2048[2], respectively. The parameters, C, h, B, and ϵ, in (4), (5), and (6) were determined empirically to be 2, 0.8, 2, and 0.5, respectively. In (10), the parameters k_1 and k_2 were determined to be 1.07 and -0.17, respectively, using a regression analysis on the human ratings for DB-2B. Fig. 8 shows the score distribution of the singing recordings in DB-2B and the resulting regression curve, i.e., (10). We can see from Fig. 8 that roughly, the smaller the value of $\mathrm{Dist}(\mathbf{O}, \mathbf{O}')$, the higher the human-rating score, and vice versa. It can also be seen from Fig. 8 that the regression curve well fits most of the data points. The root mean square error of the regression, which means the average difference between a human-rating score and system-rating score, is 2.1.

First, we introduced random errors in the note sequence of each song clip in DB-1. The resulting erroneous note sequences were then rated using (10). Here, the errors were generated by replacing the note numbers of b segments selected at random in the original sequence with random numbers between 43 and 83. A segment contained 100 consecutive frames, and the note numbers within a selected segment were replaced by the same random number. Fig. 9 shows an example of the original note sequence, along with its six error patterns generated by varying the value of b from 5 to 15, where $b = 5$ represents slight off-key, and $b = 15$ represents heavy off-key. Table II shows the results of the system rating for the erroneous note sequences, where each score was the rounded-off average of all the song

[2]Due to the limited frequency resolution, there are three notes, 44, 46, and 49, not covered in $U(j)$ of (3).

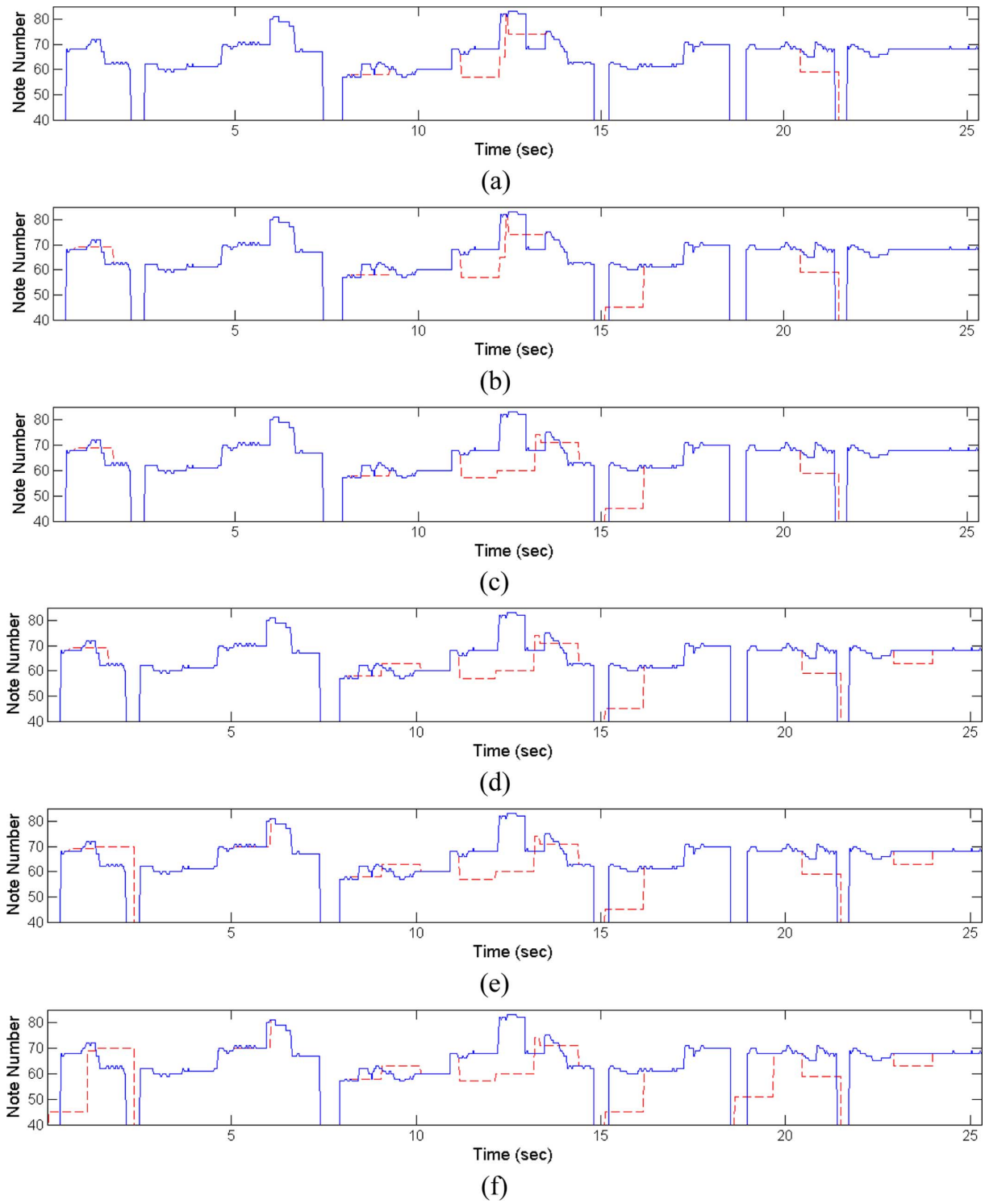

Fig. 9. Example of the original note sequence (solid line) and its six error patterns generated by varying the value of b from 5 to 15 (dotted lines). (a) $b = 5$. (b) $b = 7$. (c) $b = 9$. (d) $b = 11$. (e) $b = 13$. (f) $b = 15$.

clips' scores. We can see from Table II that the more the errors introduced, the lower the scores rated by the system.

In addition, we 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 (10). The mean and standard deviation of all the resulting scores are 32.9 and 3.6, respectively. This result 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.

Furthermore, considering that performers may add vibrato, tremolo, or wobble when singing, the resultant oscillation of pitch could introduce slight errors in our note sequence extraction. To investigate the effect of such errors on singing perfor-

TABLE II
RESULTS OF THE SYSTEM RATING FOR THE NOTE SEQUENCES OF SONG CLIPS IN DB-1 INTRODUCED WITH RANDOM ERRORS IN b SEGMENTS, IN WHICH THE REFERENCE BASES ARE THE ORIGINAL NOTE SEQUENCES WITH NO ERROR INTRODUCED

b	5	7	9	11	13	15
Score	88	73	57	52	45	30

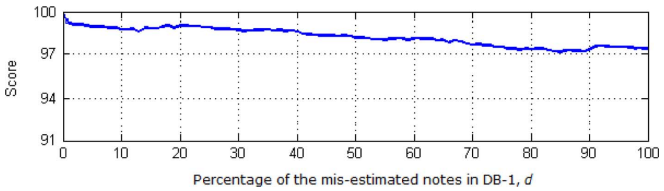

Fig. 10. Pitch-based scores with respect to the percentage of the misestimated notes in DB-1, which simulates the note extraction errors arising from the pitch oscillation in singing vibrato, tremolo, or wobble. To better show the variations, the scores were not rounded off to integers.

TABLE III
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	Score	93	90	87	70	82	80	79	75	66	69
Rating	Ranking	1	2	3	8	4	5	6	7	10	9
System	Score	83	84	80	75	78	70	70	73	61	65
Rating	Ranking	2	1	3	5	4	7	8	6	10	9

mance evaluation, we conducted an experiment by randomly shifting d percent of notes in the original note sequences of each song clip in DB-1 with ± 1 semitone and then rating the resulting note sequences using (10). Fig. 10 shows the pitch-based scores with respect to different values of d, in which the larger the value of d, the more the errors occur. We can see from Fig. 10 that there are only small differences (less than 3) between the pitch-based scores obtained before and after the errors are introduced. The results indicate that our system is insensitive to vibrato, tremolo, or wobble.

Next, experiments were conducted to rate the singing recordings in DB-2A. Table III shows the results of human rating and system rating. Each singer's score was obtained by averaging the scores of his/her 15 recordings and then rounding off to an integer. All the singers' scores were further ranked in descending order. We can see from Table III that the ranking results obtained with our system are similar to those of the human rating, though there are still significant score differences between the system rating and human rating. Overall, the system rating can well distinguish the singers in one group from another groups' singers.

2) Experiments on Volume-Based Rating: We then examined the validity of the volume-based rating using DB-2A. The parameters q_1 and q_2 in (11) were determined to be 1.12 and -0.18, respectively, using a regression analysis on the human ratings for DB-2B. Table IV shows the results of human rating and system rating. It can be seen from Table IV that the ranking results obtained with our system are roughly consistent with those of the human rating.

In addition, 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

TABLE IV
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	Score	81	90	83	74	76	79	87	84	65	68
Rating	Ranking	5	1	4	8	7	6	2	3	10	9
System	Score	83	87	82	70	73	75	80	76	62	65
Rating	Ranking	2	1	3	8	7	6	4	5	10	9

TABLE V
RESULTS OF DETERMINING THE ATTRIBUTE OF A MANUALLY-MIXED SONG CLIP TO BE "AHEAD OF A BEAT (1)", "BEHIND A BEAT (-1)", OR "IN BEAT (0)"

	K (samples)					
	Behind a beat		In beat		Ahead of a beat	
Song index	-20000	-10000	-2000	2000	10000	20000
001	-1	0	0	0	1	1
002	-1	-1	0	0	1	1
003	-1	-1	0	0	1	1
004	-1	-1	0	0	1	1
005	-1	-1	0	0	1	1
006	-1	-1	0	0	1	1
007	-1	0	0	0	1	1
008	-1	-1	0	0	1	1
009	-1	-1	0	0	0	1
010	-1	-1	0	0	1	1
011	-1	-1	0	0	1	1
012	-1	-1	0	0	1	1
013	-1	0	0	0	1	1
014	-1	-1	0	0	1	1
015	-1	-1	0	0	0	1
016	-1	-1	0	0	1	1
017	-1	-1	0	0	1	1
018	-1	0	0	0	1	1
019	-1	-1	0	0	1	1
020	-1	-1	0	0	0	1

singing recordings in DB-2A that are irrelevant to the song of the reference basis. The mean and standard deviation of all the resulting scores are 18.6 and 5.05, respectively. Such a low score indicates that the proposed volume-based rating can well recognize if a singer performs a wrong song.

3) Experiments on Rhythm-Based Rating: In the rhythm-based rating, the system uses three song-dependent HMMs to determine whether each W-length segment in a singing clip is in beat, ahead of a beat, or behind a beat. In our experiments, the numbers of states and mixture components per state used in HMMs were empirically determined to be 7 and 4, respectively. The data used for training the "out-of-beat HMMs" were generated by mixing the extracted solo singing and the accompaniment with the asynchronicity of $K = \pm 15\,000$ samples (± 0.68 sec), where "+" and "−" represent right-shifted and left-shifted of the accompaniment in the time domain, respectively.

An experiment was first conducted to investigate if the three HMMs can handle various in-beat and out-of-beat accompanied singing. The test data used here were generated by mixing the extracted solo singing and the accompaniment with the asynchronicity of $K = \pm 2000$, $\pm 10\,000$, and $\pm 20\,000$ samples. Here, the cases of $K = \pm 2000$ are perceptually in-beat, whereas the other cases are perceptually out-of-beat. Table V shows the testing results, where "1", "-1", and "0" represents that the attribute of a test segment (an entire song clip in this experiment) is determined to be "ahead of a beat", "behind a

TABLE VI
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	96	93	89	87	90	80	75	86	71	83
	Ranking	1	2	4	5	3	8	9	6	10	7

TABLE VII
OVERALL RATING BASED ON (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	74	82	77	80	79	67	72
	Ranking	1	2	3	8	4	7	5	6	10	9
System Rating	Score	85	87	84	79	81	73	70	77	63	70
	Ranking	2	1	3	5	4	7	8	6	10	9

TABLE VIII
PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT BETWEEN THE HUMAN RATING AND SYSTEM RATING

Rating Method	Correlation Coefficient
Pitch-based Rating	0.80
Volume-based Rating	0.77
Rhythm-based Rating	0.86
Overall Rating	0.82

beat", and "in beat", respectively, using (12). We can see from Table V that the attributes of most song clips are correctly determined. If "ahead of a beat" and "behind a beat" are regarded as the same attribute "out-of-beat", then the percentage of the correctly determined samples in the total test samples is 94.2%.

We then examined the validity of the rhythm-based rating using DB-2A. The length of segment, W, used in (12) was set to be 2 seconds. Table VI shows the rating results. We can see from Table VI that the ranking results obtained with our system are close to those of the human rating. This confirms the validity of the proposed rhythm-based rating.

4) Combination of Pitch-Based, Volume-Based, and Rhythm-Based Ratings: Finally, we considered the overall evaluation using (1), in which the weights w_{Pit}, w_{Vol}, and w_{Rhy} were estimated to be 0.45, 0.16, and 0.39, respectively, using the least square analysis of the human ratings for DB-2B. Table VII shows the overall rating results. It can be observed from Table VII that most of the scores obtained with the system rating match those of the human rating. Table VIII shows the Pearson product-moment correlation coefficient between human rating and system rating summarized from Tables III–VI. We can see from Table VIII that overall, there is a high positive correlation between the human rating and our system rating. This indicates the feasibility of our system in exploiting pitch, volume, rhythm-based features for singing performance evaluation.

VI. CONCLUSION

This study has developed an automatic system to assess a Karaoke singing performance. The system compares a solo singing piece with the reference Karaoke VCD music using pitch, volume, and rhythm based features. By 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 lyrics verification to further improve the system. In the context of Karaoke VCDs, there could be two ways to acquire the ground truth for lyrics verification. One is to recognize the lyrics texts in Karaoke video, and the other is to recognize the sung lyrics in Karaoke audio. Our initial study found that the former would be easier than the latter.

In addition, our future work will investigate the possibility of using regular CD music as a reference basis for singing evaluation. Given only the accompanied vocal signals available from regular CD music, an automatic singing-evaluation system may need to separate the vocals from its background accompaniment. Since there is no reliable solution at current stage to vocal extraction from regular CD music, our future work will focus on this problem.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers and the associate editors, Dr. Sylvain Marchand, for their careful reading of this paper and their constructive suggestions.

REFERENCES

[1] SingStar [Online]. Available: http://www.singstargame.com
[2] Karaoke Revolution. [Online]. Available: http://www.gamespot.com
[3] T. Nakano, M. Goto, and Y. Hiraga, "Subjective evaluation of common singing skills using the rank ordering method," in *Proc. Int. Conf. Music Percept. Cognition*, 2006.
[4] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. Int. Conf. Spoken Lang. Process. (Interspeech)*, 2006.
[5] 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 *Proc. IEEE Int. Symp. Multimedia*, 2007, pp. 75–76.
[6] P. Lal, "A comparison of singing evaluation algorithms," in *Proc. Int. Conf. Spoken Lang. Process. (Interspeech)*, 2006.
[7] O. Mayor, J. Bonada, and A. Loscos, "Performance analysis and scoring of the singing voice," in *Proc. AES 35th Int. Conf.*, 2009.
[8] J. G. Hong and U. J. Kim, "Performance Evaluator for Use in a Karaoke Apparatus," U.S. Patent No. 5,557,056, 1996.
[9] C. S. Park, "Karaoke System Capable of Scoring Singing of a Singer on Accompaniment Thereof," U.S. Patent No. 5,567,162, 1996.
[10] K. S. Park, "Performance Evaluation Method for Use in a Karaoke Apparatus," U.S. Patent No. 5,715,179, 1998.
[11] B. Pawate, "Method and System for Karaoke Scoring," U.S. Patent No. 5,719,344, 1998.
[12] T. Tanaka, "Karaoke Scoring Apparatus Analyzing Singing Voice Relative to Melody Data," U.S. Patent 5,889,224, 1999.
[13] H. M. Wang, "Scoring Device and Method for a Karaoke System," U.S. Patent No. 6,326,536, 2001.
[14] P. C. Chang, "Method and Apparatus for Karaoke Scoring," U.S. Patent No. 7,304,229, 2007.
[15] K. Omori, A. Kacker, L. M. Carroll, W. D. Riley, and S. M. Blaugrund, "Singing power ratio: Quantitative evaluation of singing voice quality," *J. Voice*, vol. 10, no. 3, pp. 228–235, 1996.
[16] W. S. Brown, H. B. Rothman, and C. M. Sapienza, "Perceptual and acoustic study of professionally trained versus untrained voices," *J. Voice*, vol. 14, no. 3, pp. 301–309, 2000.
[17] C. Watts, K. Barnes-Burroughs, J. Estis, and D. Blanton, "The singing power ratio as an objective measure of singing voice quality in untrained talented and nontalented singers," *J. Voice*, vol. 20, no. 1, pp. 82–88, 2006.

[18] G. F. Welch, C. Rush, and D. M. Howard, "Real-time visual feedback in the development of vocal pitch accuracy in singing," *Psychol. Music*, vol. 17, pp. 146–157, 1989.
[19] D. Hoppe, M. Sadakata, and P. Desain, "Development of real-time visual feedback assistance in singing training: A review," *J. Comput. Assist. Learn.*, vol. 22, pp. 308–316, 2006.
[20] A. Neri, C. Cucchiarini, H. Strik, and L. Boves, "The pedagogy-technology interface in computer assisted pronunciation training," *Comput. Assisted Lang. Learn.*, vol. 15, pp. 441–467, 2002.
[21] A. Dowd, J. J. Smith, and J. Wolfe, "Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real-time," *Lang. Speech*, vol. 41, pp. 1–20, 1998.
[22] Y. Hirata, "Computer assisted pronunciation training for native English speakers learning japanese pitch and duration contrasts," *Comput. Assisted Lang. Learn.*, vol. 17, pp. 357–376, 2004.
[23] H. M. Yu, W. H. Tsai, and H. M. Wang, "A query-by-Singing system for retrieving Karaoke music," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1626–1637, Dec. 2008.
[24] M. Piszczalski and B. A. Galler, "Predicting musical pitch from component frequency ratios," *J. Acoust. Soc. Amer.*, vol. 66, no. 3, pp. 710–720, 1979.
[25] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP)*, 1979, pp. 208–211.
[26] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov Chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, 1970.
[27] R. A. Fisher, "On the 'probable error' of a coefficient of correlation deduced from a small sample," *Metron*, vol. 1, pp. 3–32, 1921.

Wei-Ho Tsai (M'04) received the B.S. degree in electrical engineering from National Sun Yat-Sen University, Kaohsiung, Taiwan, in 1995 and the 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. He is currently an Associate Professor in the Department of Electronic Engineering and Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei, Taiwan. His research interests include spoken language processing and music information retrieval.

Hsin-Chieh Lee received the B.S. degree in electronic engineering and the M.S. degree in computer and communication engineering from National Taipei University of Technology, Taipei, Taiwan, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree in computer and communication engineering at National Taipei University of Technology. His research interests include signal processing and multimedia applications.