AN AUTOMATED SINGING EVALUATION METHOD FOR KARAOKE SYSTEMS

Wei-Ho Tsai and Hsin-Chieh Lee

Department of Electronic Engineering & Graduate Institute of Computer and Communication Engineering, National Taipei University of Technology, Taipei, Taiwan
{whtsai,t9419004}@ntut.edu.tw

ABSTRACT

Although many Karaoke systems come with a scoring feature that provides singers with a simple performance rating, the automated scoring, however, is either random or incomparable to that of human evaluation. 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 VCD music for the training of our proposed system. Our experiment shows that the results of automated singing evaluation are close to the human rating.

Index Terms—Accompaniment, Karaoke, Singing Evaluation

1. INTRODUCTION

Karaoke is a popular recreational pastime that allows people to 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, Karaoke is a convenient way to help people practice singing. Thus, a Karaoke machine with an intelligent singing evaluation capability is useful to provide singers with an immediate feedback.

However, although many Karaoke apparatuses come with an automated scoring feature, their evaluation capabilities do not always match that of human evaluation. A cause of blunder arises from the fact that an automated 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.

Recently, a few studies have been made to improve the capability of automated singing evaluation. Nakano *et al*. [1] design a 2-class (good/poor) classifier based on support vector machine 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*. [2] develop a singing skill visualization interface, which analyzes and visualizes the pitch contours of a test singing sample and the vocal-part in music CD recordings. Lal [3] proposes two pitch-based similarity measures to determine how close a user's singing clip is to the reference singing (a cappella) clip. In [4], 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-related cues and present a preliminary experiment results.

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 different levels of singing capabilities for recordings of solo Karaoke performances. The performances were rated independently by four musicians, and then used in conjunction with additional Karaoke VCD music to train our proposed system. Our experiment shows that the results of automated singing evaluation are close to the human rating.

2. SYSTEM OVERVIEW

Given a Karaoke singing performance, the aim of our system is to intelligently evaluate the performance such that the scoring strongly correlate to the subjective judgments of musicians. To match the rating with that of human evaluation, a reference basis is required for the system. In this study, the reference basis is the vocal part extracted from 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.

Fig. 1 shows the proposed singing-evaluation system. When a singing piece is evaluated, the system performs pitch-based analysis, volume-based analysis, and rhythm-based analysis. The resulting scores from each component are then combined according to a heuristic rule: $0.5 \cdot S_{\text{Pit}} + 0.2 \cdot S_{\text{Vol}} + 0.3 \cdot S_{\text{Rhy}}$.

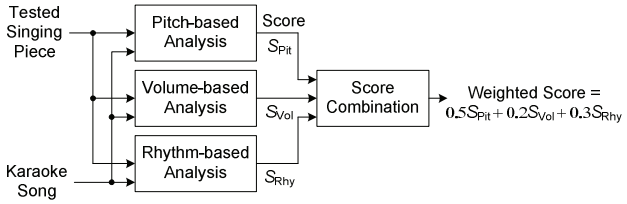

Fig. 1. The proposed singing-evaluation system.

3. PITCH-BASED ANALYSIS

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 analysis begins by converting the waveform of a singing recording into a sequence of MIDI notes. Let n_1, n_2,…, n_M be the inventory of possible notes performed by a singer. Our aim is to determine which among the M possible notes is most likely sung at each instant. We apply the method in [5] to solve this problem. First, the vocal signal is divided into frames by using a fixed-length sliding Hamming window. 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 $1 \le j \le J$. Then, the signal's energy on m-th note in frame t can be estimated by

$$\hat{x}_{t,m} = \max_{\forall j, U(j)=n_m} x_{t,j} , \tag{1}$$

and

$$U(j) = \left\lfloor 12 \cdot \log_2\left(\frac{F(j)}{440}\right) + 69.5 \right\rfloor, \tag{2}$$

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. Next, we use the strategy of Sub-Harmonic Summation (SHS) [6] to estimate the sung notes. It computes the "strength" of note e_m in frame t using:

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

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, i.e.,

$$o_t = \arg\max_{1 \le m \le M} y_{t,m} . \tag{4}$$

The resulting note sequence could be refined by taking into account the continuity between frames, since a note usually lasts several frames. We use median filtering, which replaces each note with the local median of notes of its neighboring frames, to remove jitters between adjacent 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 VCD, 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 [7].

Fig. 2 shows the block diagram of the pitch-based analysis. In the off-line phase, a Karaoke song's accompanied vocal signal is converted from its waveform representation $v[n]$ into a reference note sequence $\mathbf{O} = \{o_1,o_2,...,o_T\}$. 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 on-line phase, a singing recording is converted from its waveform signal $s'[n]$ into a note sequence $\mathbf{O}' = \{o'_1,o'_2,...,o'_{T'}\}$. Then, the performer's singing skill is evaluated on the basis of the similarity 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 $\mathbf{O}'$.

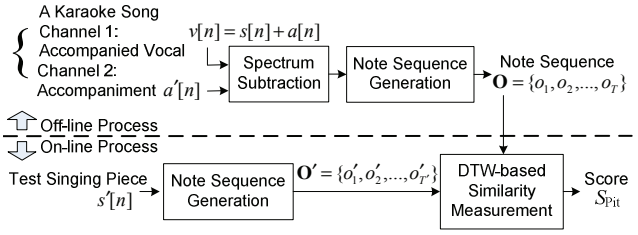

Fig. 2. Pitch-based analysis.

DTW constructs a $T \times T'$ distance matrix $\mathrm{D} = [D(t,t')]_{T \times T'}$, 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:

$$D(t,t') = \min \begin{cases} D(t-2,t'-1) + 2 \times d(t,t') \\ D(t-1,t'-1) + d(t,t') - \varepsilon \\ D(t-1,t'-2) + d(t,t') \end{cases} , \tag{5}$$

and

$$d(t,t') = | \, o_t - o_{t'} \, | , \tag{6}$$

where ε is a small constant that favors the mapping between notes o_t and $o_{t'}$, given the distance between note sequences $\{o_1,o_2,...,o_{t-1}\}$ and $\{o_1,o_2,...,o_{t'-1}\}$. After the distance matrix $\mathbf{D}$ is constructed, the similarity between $\mathbf{O}$ and $\mathbf{O}'$ can be evaluated by

$$\mathrm{Sim}(\mathbf{O},\mathbf{O}') = \begin{cases} \max_{T/2 \le t' \le \min(2T,T')} \left[-D(T,t') \right], & \text{if } \frac{T}{2} \le T' \le 2T \\ -\infty , & \text{otherwise} \end{cases} , \tag{7}$$

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

$$S_{\mathrm{Pit}} = 100 \cdot k_1 \exp[k_2 \cdot \mathrm{Sim}(\mathbf{O},\mathbf{O}')/T], \tag{8}$$

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

4. VOLUME-BASED ANALYSIS

When a song is composed, abbreviations called dynamics are notated in music scores to indicate the volume when performing the song, 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. We characterize the volume variations by a sequence of frame energies.

Fig. 3 shows the processes of the volume-based analysis. As Karaoke VCD music does not contain solo singing, direct comparison of energy sequences between a test singing and its reference singing is infeasible. To solve this problem, we estimate the 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 analysis, we apply DTW to measure the similarity Sim($\mathbf{E}$,$\mathbf{E}'$) between energy sequence of the reference singing, $\mathbf{E}$, and sequence of the test singing, $\mathbf{E}'$. Then, a volume-based score is obtained using

$$S_{\text{Vol}} = 100 \cdot q_1 \exp[q_2 \cdot \text{Sim}(\mathbf{E}, \mathbf{E}')/T], \tag{9}$$

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

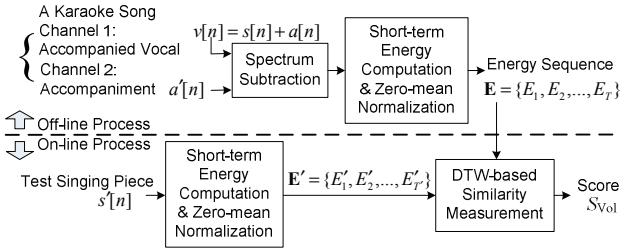

Fig. 3. Volume-based analysis.

5. RHYTHM-BASED ANALYSIS

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 pre-recorded, the singer must follow the pace of the accompaniment; otherwise, the performance may sound out of beat. Thus, our basic strategy is to evaluate 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. 4 shows the block diagram of the proposed rhythm-based analysis. The basic strategy is to represent synchronous and asynchronous accompanied singing by probabilistic models and then perform stochastic recognition. For each song, a "synchronous model" is built using the accompanied vocal signal extracted from Karaoke VCD music. To capture the temporal variations of the signal's harmonic structures, we convert the signal's waveform into a sequence of strength vectors using Eq. (3), and then represent it by a hidden Markov model (HMM). Specifically, the observations for the HMM are a sequence of vectors $\{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_t, ...\}$, where $\mathbf{Y}_t = [y_{t,1}, y_{t,2}, ..., y_{t,M}]^{\text{Tr}}$, and $y_{t,m}$, $1 \le m \le M$, is the strength of note e_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 [8]. We denote the synchronous HMM by λ_1.

As to synchronous accompanied singing, we create two 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 like 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 Eq. (3). The sequence is then represented by an HMM using Baum-Welch algorithm. We denote this HMM by λ_2. 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 λ_3.

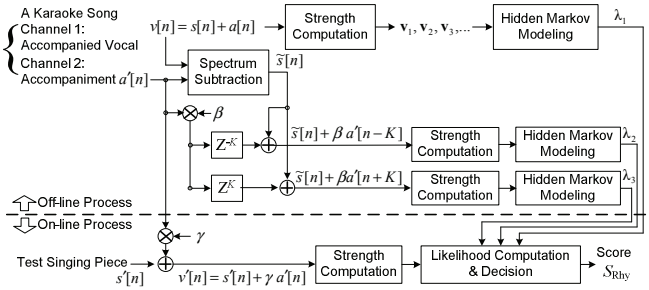

Fig. 4. Rhythm-based analysis.

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 $\mathbf{G}_1$, $\mathbf{G}_2$,..., $\mathbf{G}_L$. Next, the attribute of each segment is determined by

$$A_\ell = \arg\max_{1 \le j \le 3} \Pr(\mathbf{G}_\ell \mid \lambda_j), \tag{10}$$

where A_ℓ = 1, 2, and 3 represent that the singing is in beat, ahead of a beat, and behind a beat, respectively. As (A_ℓ = 2,3) indicates the occurrence of incorrect rhythm, the system computes a rhythm-based score using

$$S_{\text{Rhy}} = 100 \cdot \frac{1}{L} \sum_{\ell=1}^{L} \delta(A_\ell - 1), \tag{11}$$

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

6. EXPERIMENTS

6.1. Music Data

Our music data consists of two databases. The first database, denoted as DB-1, contains 20 Mandarin tracks extracted from Karaoke VCDs. For computational efficiency, each extracted music track was downsampled from 44.1kHz to 22.05kHz and stored as PCM wave. The second database contains songs recorded by ourselves in a quiet room. We employed 25 singers to record for solo vocal samples. Each singer performed one verse/chorus part of each of the 20 Mandarin songs. The recordings range in

duration from 25 to 40 seconds. They 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 them 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 often cannot follow the tune, and some of tem even never sing Karaoke before. We marked them by Group III. In addition, to establish the ground truth for automated singing evaluation, we employed four musicians to rate the singing recordings independently. The rating results given by the four musicians were then averaged to form a reference score for each singing recording.

6.2. Experiment Results

We divided DB-2 into two subsets. The first one, denoted by DB-2A, was used to test our system. It contains 150 recordings = (10 singers) × (first 15 recordings/singer). Among the 10 singers in DB-2A, 2 are 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 Eqs. (8) and (9). It contains the remaining recordings not covered in DB-2A. The numbers of states and mixture components per state used in HMMs were empirically determined to be 7 and 4, respectively.

Table 1 shows the singing-evaluation results for DB-2A. Here, we compared 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 1 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. The results confirm the feasibility of our singing-evaluation system.

7. CONCLUSION

This study has developed an automated 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 automated 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.

8. ACKNOWLEDGEMENT

This work was supported in part by the National Science Council, Taiwan, under Grant No. NSC 99-2628-E-027-005.

9. REFERENCES

[1] T. Nakano, M. Goto, Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," *Interspeech*, 2006.

[2] T. Nakano, M. Goto, Y. Hiraga, "MiruSinger: a singing skill visualization interface using real-time feedback and music CD recordings as referential data," *IEEE Int. Symp. Multimedia*, 2007.

[3] P. Lal, "A comparison of singing evaluation algorithms," *Interspeech*, 2006.

[4] O. Mayor, J. Bonada, and A. Loscos, "Performance analysis and scoring of the singing voice," *AES 35th International Conference*, 2009.

[5] 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, 2008.

[6] 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.

[7] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *ICASSP*, 1979, pp. 208-211.

[8] 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.

Table 1. Singing-evaluation Results for the 10 singers in DB-2A.

(a) Pitch-based Evaluation

Singer Index		A	B	C	D	E	F	G	H	I	J
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	82	84	81	77	78	69	70	74	62	68
Rating	Ranking	2	1	3	5	4	8	7	6	10	9

(b) Volume-based Evaluation

Singer Index		A	B	C	D	E	F	G	H	I	J
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

(c) Rhythm-based Evaluation

Singer Index		A	B	C	D	E	F	G	H	I	J
Group		I	I	II	II	II	II	II	II	III	III
Human	Score	90	87	83	80	87	72	77	81	70	79
Rating	Ranking	1	2	4	6	3	9	8	5	10	7
System	Score	96	93	89	87	90	80	75	86	71	83
Rating	Ranking	1	2	4	5	3	8	9	6	10	7

(d) Overall Evaluation (Combination of Pitch-based, Volume-based, and Rhythm-based Evaluation)

Singer Index		A	B	C	D	E	F	G	H	I	J
Group		I	I	II	II	II	II	II	II	III	III
Human	Score	90	89	85	74	82	77	80	79	67	72
Rating	Ranking	1	2	3	8	4	7	5	6	10	9
System	Score	86	87	84	79	81	74	74	78	65	72
Rating	Ranking	2	1	3	5	4	7	8	6	10	9