An Automatic Grading Method for Singing Evaluation

Zeyu Jin, Yuxiang Liu, Jia Jia, Lianhong Cai
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology,
Tsinghua University, Beijing, China

***Abstract*—Automatic evaluation of singing quality is essential to singing practitioners. In this paper, we present a methodology that generates grades and comments for singing quality. In this method, we first perform note segmentation and pitch estimation. Then a method based on experience of human perception is used to evaluate the accuracies of intonation and rhythm. A subjective evaluation is carried out in which spearman's correlation is used to evaluate the consistency between the proposed method and experts' opinion. The result demonstrates high reliability of the proposed annotation system and grading method.**

Keywords-component; automatic singing evluation, grading method, rank ordering

I. INTRODUCTION

Singing quality evaluation is important for anyone who wants to improve singing skills. Nowadays however, reliable evaluation is dependent on experts and existing automatic evaluating systems cannot achieve the same objectivity in consistency with the experts. The difficulties come in two aspects: First, pitch estimation and note segmentation are hard to achieve. Second, a sophisticated method is necessary for grading so that evaluation result achieved from the "purely-objective" data comes in consistence with experts' subjective opinion. The term "reliable" is defined to meet the requirement that these two problems are solved properly. Our work is to design such a system for automatic singing evaluation where singers can change tempo and key arbitrarily. Our methods do not require accompany or other additional information but only a piece of singing recording and its reference score.

The framework of our method is as follows:

First, we perform automatic annotation in which a piece of singing voice is separated into notes. We also implemented algorithms to determine a precise onset of each note so that the evaluation on rhythm accuracy can be reliably achieved.

Second, we estimate pitch and length for each note. If the singing voice involves big errors such as continuous intonation deviation, the segmentation process will fail which will result in a penalty. Otherwise, in-depth information will be provided for each note that a user sung. This information is important for practitioner to identify his or her problems in singing.

This work is supported by National Natural Science Foundation of China (60805008, 61003094, 60928005, 60910130).

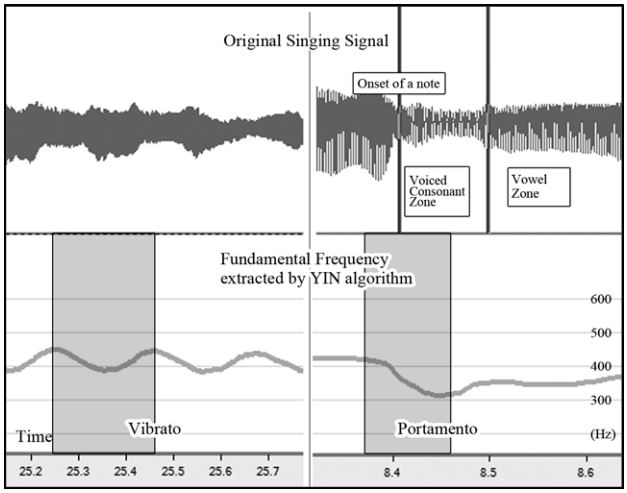
Finally, we summarize grading result for each note and provide overall grading for intonation and rhythm respectively. By comparing this automatic grading with experts' evaluation result, the reliability of our system is assessed.

II. RELATED WORK

To obtain musical information from audio signal, a number of studies have been performed on fundamental frequency and score-to-audio alignment. Achievement is seen in the domain of instrumental music but still sparse for singing voice. Alain de Cheveigne presented YIN, a fundamental frequency estimator with relatively low error-rates [1]. To align audio frames to notes, Orio and Schwarz proposed a new methodology for automatic alignment based on dynamic time warping [2]. More recently, Liu et al. presented IMED (The Intelligent Music Editor) for automatically aligning, analyzing and editing over multi-track recordings [3]. The essence of these methods is employed in our system and adjusted so as to satisfy the characteristics of singing voice.

There are two forms of study on automatic evaluation for singing quality. The first focuses on the quality of voice such as [4] where timbre, brightness, vibrato and other expressiveness are the major concern of the study. Another form of evaluation concerns the accuracy of pitch and rhythm, where musical information extracted from singing signal is critical for the evaluation and grading. For most automatic grading systems such as [5] and [6], their scoring rules are often too simple to be reliable especially when applied in our case where singing tempo and key are arbitrarily chosen by the singer. To determine pitch for each singing segment, merely calculating the average F0 may deviate severely from human perception especially when vibrato and portamento are associated (Fig. 1). In the real case, adding these expressive techniques does not affect people's perception of pitch. Our grading algorithm not only corresponds with acoustic characteristics of singing signals, but also in accord with expertise evaluation over professional singing features.

In this paper, we first briefly introduce a method that automatically annotates fundamental frequency and segmentation based on audio-to-score alignment. Then we concentrate on the algorithm that evaluates singing quality over the data achieved by the annotation phase. To present the reliability of those methods, we show a subjective evaluation on experts. We compare the experts' judgment on



(a) (b)

Figure 1. Viberatro and Portamento. (a) A typical sample of vibrato; the F0s repeatedly swing over the percepted freqency. (b) A typical sample of Portamento: Slow gliding F0 curve appears at the end of the last singing note and the start of the current one; this phenamenon does not affect the percepted pitch but makes it complicated to determine the accurate pitch automatically.

singing clips with our system's to assess the consistency between the automatic technique and human judgment.

III. AUTOMATIC ANNOTATION

This procedure is to annotate elements on audio sequence including F0 for each frames, onset and ending of notes and their vowel and consonant part. We distinguish these elements especially consonant because precise onset of each note is determined by analyzing the attributes of the frames. A brief discretion of our annotation method is in below.

- Step1: Estimating Fundamental Frequency (F0). In this step, YIN algorithm is employed with some modification done upon. First, we calculate the average magnitude difference between a segment of signal and another segment lagged by a trail period with a same length so that an AMDF curve is generated. Then we search throughout the curve for valleys and uses parabola estimation to determine the accurate time of the valley. However, this method suffers octave errors and to prevent it, the three most likely candidates of valleys are preserved and F0 is selected according to its adjacent frames. The error rate is several times lower than the original YIN algorithm. Meanwhile, we also preserve the AMDF rate of the selected valley. This information is important to determine the F0 reliability of the frame.
- Step 2: Annotate the "unreliable frames". For the frames whose F0 has obvious octave error and frames without F0 such as voiceless consonant and unvoiced area, this step identifies them and annotates them as unreliable ones. The F0 data in reliable frames are most important in the key transportation (Step 3) and coarse alignment (Step 4). For each unreliable frame, we predict their properties (voiceless consonant, octave error or unvoiced area) by checking the frames amplitude, AMDF rate and zero-crossing coefficients.
- Step 3: Determine the key transportation. We compute DTW alignment over the reliable frames and the score. The difference between a frame pitch and score is defined as such:

$$D(x, y) = (F0(x) - F_s(y))^2 \qquad (1)$$

where, F0 corresponds the fundamental frequency of the frame and f_s indicates the corresponding frequency of the note in the score. We compute this test repeatedly for each key-transportation. The test with minimum DTW value indicates that its corresponding key transportation best suits the audio. Information about DTW can be found in [2].

- Step 4: Coarse Alignment. In this phase, we conduct DTW audio-to-score alignment using a new method where we not only uses reliable frames but also employs the unreliable frames especially those identified as voiceless consonant as indicators of an onset. In DTW, we compare both onsets and pitch difference so that more accurate result is achieved.
- Step 5: Refinement. Since consonant information is achieved in the second step, we move the onset towards its adjacent beginning of a continuous area dominated by voiceless consonant. Then we correct note ending that has an unvoiced area before it and do other refinement to improve segmentation accuracy. Till now, the annotation phase is done completely.

Over our previous testing data where intonation accuracy of the recording is within a whole tone, the alignment accuracy is achieved with an average error of approximately 24ms (6 frames). For recordings with continuous intonation error accompanied with unstable tempo change, the annotation process may fail. So our system will automatically identify whether the alignment phase is done successfully and if not, the grading phase is terminated and a low score is provided to the singer indicating great errors have occurred in his or her singing.

IV. GRADING METHOD

This step includes grading for each note on each parameters and the grading for each parameters over all the notes. Just like the parameters used in [7], we also use Intonation Accuracy and Rhythm Accuracy for our grading. Their definition is as follows:

- Intonation Accuracy: The accuracy of the relative pitch, appropriate key transposition is allowed in this criterion.
- Rhythm Accuracy: The stability of tempo.

A. Intonation Accuracy

Based on the segmentation result of the last step, we can evaluate pitch for each note. It is hasty to calculate the average pitch of the segment without proper adjustment

because vocal elements (such as voiceless consonant) and singing techniques (such as vibrato and portamento (Fig. 1)) may affect the consistency between the average pitch and a heard pitch. So our method is proposed to solve these problems. The steps are as follows:

- Step 1: Determine reliable pitch area. In this step, we first eliminate the frames that are labeled "unreliable" in the automatic annotation so that octave error and voiceless consonant are excluded with the beginning of the area shifted forward. Then we truncate the onset and ending of the segmentation about 40ms to cut off possible portamento. The remained frames form the reliable pitch area.
- Step 2: Construct the histogram of fundamental frequency of the reliable pitch area.
- Step 3: Cut the 30% of edge of the histogram and calculate the average of the remained.

The intonation deviation for each note is defined as such:

$$I(x) = \begin{cases} 1, & x = 1 \\ \left| p^{(x)} - p^{(x-1)} - (p_0^{(x)} - p_0^{(x-1)}) \right|, & x = 2,3,...,N \end{cases} \quad (2)$$

$$p^{(x)} = 12 \log_2 \frac{F_0^{(x)}}{F_{5C}} + 60 \quad (3)$$

I(x) stands for intonation accuracy for the x-th note, N donates the number of notes and $p_0^{(x)}$ and $p^{(x)}$ means the x-th note's actual and sung pitch in the form of MIDI code for pitch (where 60 stands for Tenor C and 1 points donates a semitone). This formula calculates the error rate of chromatic interval between two adjacent notes. We do not use absolute pitch since we allow for key transportation.

To grade intonation accuracy over the deviation of notes, we design a function (4) called Bounded Deviation Grade (BDG) that is used in grading both the two aspects.

$$BDG(d, k_1, k_2, \alpha) = \begin{cases} 1, & , I(x) \le t_1 \\ 1 - \left(\frac{d - t_1}{t_2 - t_1} \right)^{\alpha} & , I(x) > t_1, I(x) < t_2 \\ 0 & , I(x) \ge t_2 \end{cases} \quad (4)$$

In this function, k_1 and k_2 are the lower bound of a full grade and the upper bound of a zero grade. α is used to indicate the grading curve between bound k_1 and k_2. A higher α can make larger errors more distinguishable among smaller errors while it renders small errors' grade close to each other. The result of the function ranges from 0 to 1 which defines a zero grade and a full grade.

The intonation accuracy is calculated as such:

$$IA(x) = BDG(I(x), k_1, k_2, \alpha) \quad (5)$$

The overall Intonation accuracy is achieved by calculating the average of $IA(x)$.

In our system, k_1=0.2 and k_2=2.0 are used. Since we use ranking method in subjective evaluation which is not affected by the selection of we set it to 1.0 for convenience.

B. Rhythm Accuracy

Based on automatic annotation, segmentation of notes over the singing voice can be directly used to indicate the rhythm of the recording. The tempo of each note can be achieved by applying (6)

$$tempo_n = \frac{onset_{n+1} - onset_n}{loc_{n+1} - loc_n} \quad (6)$$

where $onset_n$ donates the audio onset of the n-th note and loc_n donates the note's location in score, where we typically use 1.0 to present the length of a quarter note. To trace the tempo changes in the recording, we define a parameter called LTR (lagged tempo reference) to simulate human perception in evaluating tempo accuracy. We consider that the experts have a reference tempo in mind and if the actual tempo deviates from the referred tempo, an error is recognized. It is possible that the singer stabilize his or her reference tempo after the heard tempo changes and the listener's refernece tempo will also follow the singer's after stable tempo is recognized. LTR is such a parameter that estimates the listener's reference tempo. Its definition is as follows.

$$LTR(n) = LTR(n-1) \times lag + tempo_n \times (1 - lag) \quad (7)$$

where the variable *lag* is a parameter that measures how fast LTR reference tempo follows the heard tempo. The error of the singing note is simply the subtraction of the actual tempo and the LTR referring tempo. We define the normalized tempo deviation of n-th note in (8) and use BDG function to generate its grade (9).

$$d(n) = \frac{\left| LTR(n) - tempo_n \right|}{\sqrt{LTR(n)^2 + tempo_n^2}} \quad (8)$$

$$RA(x) = BDG(d(n), k_1, k_2, \alpha) \quad (9)$$

In our subjective evaluation, we consider a doubled tempo change $|\log_2(tempo_n - LTR(n))| = 1$ as zero grade error and thus $k_2 = 1/\sqrt{5} = 0.447$. We set k_1=0, α=1 and the average of $RA(x)$ as the overall rhythm accuracy.

V. EXPERIMENT AND DISCUSSION

A. Subjective Evaluation

The aim of subjective evaluation is to test the system's consistency with human judges on the two aspects, intonation and rhythm accuracy. Since grading criteria vary among human judges, we use rank ordering method instead of grading in numbers just like the work of [7] and [8]. Seven songs are used in the experiment. For each song, six singing recordings are provided which were sung by three male and three female singers. For convenience, the singers are asked to sing in pitch names so that when error occurs, the human judges will easily make judgment about which note is sung improperly.

The subjective evaluation is conducted by providing each judge a set of files including an instruction sheet and seven the singing clips. Ten experts with solid skills of listening accompanied with background of professional singing, musical instruments or conducting are invited in this experiment. They are asked to give grade for each song and then sort them into ranks. For songs with the same grade, the judges need to either give them a rank forcefully or give them a rank number which is the average the rank number of the tied samples (For example, two samples are tied in rank 3, their rank number will be both 3.5). In this way, random error caused by samples with very close quality is reduced, which also releases the stress of scoring samples with similar quality. In the experiment, we allow all experts give scores to the samples in their own view. After the evaluation, the experts are required to write down their selected criteria which can be used in further study.

To measure the ranking consistency between our system and those of the experts, we use the Spearman's rank correlation coefficients ρ [9] as is defined in (10):

$$\rho = 1 - \frac{6}{N^3 - N}\sum_{i=1}^{N}(a_i - b_i)^2 \qquad (10)$$

Where N (=6) is the number of recordings for each singing clips, and a_i b_i are the i-th rank value of human judgment and automatic evaluation. The value of ρ ranges from -1 (a reversed order) to 1 (a same order). In the experiment we use this equation to evaluate the consistency between the automatic method and the experts' and the inter-consistency among the experts.

B. Results and Discussions

TABLE I. SUBJECTIVE EVALUATION

Clip	Subject	Average ρ of our method	Average ρ of experts
1	Intonation	0.7643	0.8384
	Rhythm	0.8600	0.8676
2	Intonation	0.7779	0.6585
	Rhythm	0.8679	0.8068
3	Intonation	0.7571	0.7359
	Rhythm	0.8000	0.6500
4	Intonation	0.8321	0.8068
	Rhythm	0.9171	0.9006
5	Intonation	0.7243	0.6616
	Rhythm	0.4857	0.5094
6	Intonation	0.7543	0.6959
	Rhythm	0.4243	0.4952
7	Intonation	0.8321	0.7989
	Rhythm	0.6629	0.6803
overall	Intonation	0.7774	0.7423
	Rhythm	0.7168	0.7014
Overall Average		0.7471	0.7219

The ranking correlation is calculated between the automatic ranking and the subject ranking, and is also calculated among pairs of subject ranking. Table 1 shows both the average ρ of our system and the average ρ of the experts for each singing clip. Fig. 2 shows the comparison between the average ρ of our system and the average ρ of the experts. For certain singing clips there is conflict among experts which will surely affect the results obtained by our

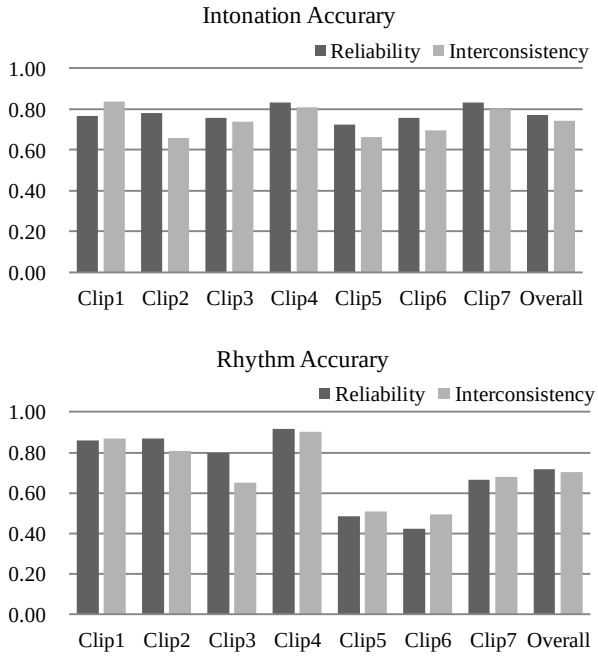


Figure 2. Correlation between system reliability (average ρ) and Inter-consistency among experts.

system. Clips with higher inter-consistency among experts will be more important in evaluating our system. And if our system scores more than the average correlation rate among the experts, we can imply that our system is more reliable than the average of the experts.

The results show us that for clips with higher inter-consistency, the system reliability is significantly higher. The overall spearman's correlation rate is 0.7774 for intonation and 0.7168 for rhythm. And for most of the clips, our system's consistency rate is larger than the inter-consistency of the judges.

VI. Conclusion and Future Work

In this paper, we describe a grading method for singing evaluation on the aspects of intonation accuracy and rhythm accuracy. We propose a framework of reliable singing evaluation which consists of an automatic annotation system and our grading algorithm. The experiment on subjective evaluation demonstrates strong consistency between our method and human perception and thus our proposed system and methods are reliable and applicable.

According to the subjective criteria collected from the subjects, problem still remains that no common conceptual criteria for singing evaluation is accepted among the experts. Several useable criteria are demonstrated in the works like [7] and [8], but are not strong enough to be the prototype for designing automatic methods. Further studies will look for stronger criteria that better consist with human perception.

To obtain more reliable evaluation result based on our proposed framework, we also need to improve accuracy for note segmentation and pitch estimation. For singing sample with severe singing errors, our annotation system fails to recognize accurate notes from the wrong ones while the experts can. Future work will also involve estimating expert's listening skills to recognize and categorize all discernable singing errors.

References

[1] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music", The journal of the Acoustical Society of America 111, No.4, 2002

[2] N. Orio, and D. Schwarz, "Alignment of Monophonic and Polyphonic Music to a Score", Proceedings of the ICMC 2001, Havana, Cuba, October 12, 2001

[3] Y. Liu, R. B. Dannenberg, and L. Cai, "The Intelligent Music Editor: Towards an Automated Platform for Music Analysis and Editing", ICIC 2010

[4] C. Watts, K. Barnes-Burroughs, J. Estis, D. Blanton, "The Singing Power Ratio as an Objective Measure of Singing Voice Quality in Untrained Talented and Nontalented Singers", Journal of Voice, Volum 20, Issue 1, pp. 82-88, March 2006

[5] S. Kang and J. Park, "System and method for grading singing data", WO Patent WO/2006/115,387, 2006

[6] T. Nakano, M.Goto, and Y. Hiraga, "An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitched Interval Accuracy and Vibrato Features", Ninth International Conference on Splken Language Processing, 2006

[7] C. Cao, M.Li, Jian. Liu and Y. Yan, "A Study on Singing Performance Evaluation Criteria for Untrained Singers", ISCP Proceedings 2008

[8] T. Nakano, M. Goto, Y. Hiraga, "Subjective evaluation of common singing sklls using the rank ordering method", 9th Interantional Conference on Music Perception and Cognition, 2006

[9] M. Kendall and J. D. Gibbons "Rank Correlation Methods", New York: Oxford University Press, 1990