An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features

Tomoyasu Nakano  
Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan  
nakano@slis.tsukuba.ac.jp

Masataka Goto  
National Institute of Advanced Industrial Science and Technology (AIST), Japan  
m.goto@aist.go.jp

Yuzuru Hiraga  
Graduate School of Library, Information and Media Studies, University of Tsukuba, Japan  
hiraga@slis.tsukuba.ac.jp

Abstract

This paper presents a method of evaluating singing skills that does not require score information of the sung melody. This requires an approach that is different from existing systems, such as those currently used for Karaoke systems. Previous research on singing evaluation has focused on analyzing the characteristics of singing voice, but were not aimed at developing an automatic evaluation method. The approach presented in this study uses pitch interval accuracy and vibrato as acoustic features which are independent from specific characteristics of the singer or melody. The approach was tested by a 2-class (good/poor) classification test with 600 song sequences, and achieved an average classification rate of 83.5%.

Index Terms: singing skill, automatic evaluation, unknown melodies.

1. Introduction

The aim of this study is to explore a method of automatic evaluation of singing skills without score information. Our interest lies in identifying the criteria that human subjects use in judging the quality of singing for unknown melodies, using acoustic features which are independent from specific characteristics of the singer or melody. Such evaluation systems can be useful tools for improving singing skills, and also can be applied to broadening the scope of music information retrieval and singing voice synthesis.

Previous work related to singing skills include those based on a control model of fundamental frequency (F0) trajectory [1], general characteristics [2, 3], as well as work on automatic discrimination of singing and speaking voices [4], and acoustic differences between trained and untrained singers’ voices [5, 6, 7]. None of these work have gone as far as presenting an automatic evaluation method.

This paper presents a singing skill evaluation scheme based on pitch interval accuracy and vibrato, which are regarded as features that function independently from the individual characteristics of singer or melody. To test the validity of these features, an experiment of automatic evaluation of singing performance by a 2-class classification (good/poor) was conducted.

The following sections describe our approach and the experimental results of its evaluation. Section 2 presents discussion of features. Section 3 describes the classification experiment and its evaluation. Section 4 concludes the paper, with discussion on directions for future work.

2. Acoustic Features for Automatic Singing Skill Evaluation

Human subjects can be seen to consistently evaluate the singing skills for unknown melodies [8]. This suggests that their evaluation utilizes easily discernible features which are independent of the particular singer or melody. Within the scope of this paper, we propose pitch interval accuracy and vibrato as such feature candidates.

The proposed features are numbered and shown in boxed form, like \( \text{[boxed number]} \). Throughout the paper, the singing samples are solo vocal and are digital recordings of 16bit/16kHz/monaural.

2.1. Estimating Pitch Interval Accuracy

The pitch interval accuracy is judged by the fitting of the F0 (fundamental frequency) trajectory to a semitone (100 cent) width grid (corresponding to equal temperament in the Western Music Tradition). Hereafter, pitch/frequency values will be referred to by cents, which are log-scale frequency values. The cent value \( f_{\text{cont}} \) of frequency \( f_{\text{fit}} \) given as follows (middle C corresponds to 4800 cent).

\[
f_{\text{cont}} = 1200 \log_2 \frac{f_{\text{fit}}}{440 \times 2^{1/5}} \quad (1)
\]

Suppose the semitone grid borders are set at multiples of 100. Then a particular pitch \( x \) has an offset of \( F(0 \leq F < 100) \) from its nearest lower border. If \( F \) has a constant value throughout the song sequence, then the singing can be seen to have a good fitting to the semitone grid.

Let \( p(x; F) \) be a Gaussian comb filter for pitch \( x \) and offset \( F \) defined as follows, where \( \omega_i \) is a weight factor (currently set to 1), and \( \sigma_i \) (16 cent) is the standard deviation of the Gaussian distribution.

\[
p(x; F) = \sum_{i=0}^{\infty} \frac{\omega_i}{\sqrt{2\pi} \sigma_i} \exp \left\{ -\left( \frac{x - F - 100i}{2 \sigma_i^2} \right)^2 \right\} \quad (2)
\]

Using this as a filter function, the semitone stability \( P_g(F, t) \) at time \( t \) and width \( T_g \) is defined as follows, where \( F_{\text{F0}}(t) \) denotes the F0 and \( P_{F0}(t) \) denotes the F0 possibility at time \( t \), which are estimated per 10 msec using the method of Goto et al. [9].

\[
P_g(F, t) = \int_{t-T_g}^{t} p(F_{\text{F0}}(\tau); F) P_{F0}(\tau) d\tau \quad (3)
\]

If all the F0 values are quantized by units of 100 cent, \( P_g(F, t) \) will have a single sharp peak at grid frequency \( F_g \). The sharpness
of the \( P_s(F, t) \) distribution thus indicates the degree of deviation of the singer from any semitone grid.

Figure 1 shows the calculation process and example results of \( P_s(F, t) \). The top figure shows the original F0 trajectory. The second figure shows the result of smoothing by an FIR lowpass filter with 5 Hz cutoff frequency\(^1\), and then removing the silent sections. The purpose of lowpass filtering is to remove the F0 fluctuations (overshoot, vibrato, and preparation) of the singing voice [1]. \( P_s(F, t) \) is calculated with a 200-sample (2 sec) rectangular window shifted by 5 samples. The bottom left figures show example fitting to grids with \( F = 19 \) and \( F = 78 \), which are cumulated as \( P_s(F, t) \) values shown to the right.

Figure 2 shows examples of \( P_s(F, t) \) and its long-term average \( g(F) \). When the singing is “good”, then \( P_s(F, t) \), and consequently, its \( g(F) \) always has a single sharp peak. So the sharpness of \( g(F) \) can be utilized as a measure of pitch interval accuracy. Its second moment \( M \) (defined as follows) is used as feature 1:

\[
M = \int_{F_g-50}^{F_g+50} (F - F_g)^2 g(F) dF
\]

(4)

where \( F_g \) is the \( F \) value for maximum \( g(F) \), i.e.:

\[
F_g = \arg \max_F g(F)
\]

(5)

The second feature \( 2 \) is taken as the slope \( b_g \) of the linear regression line of function \( G(F) \):

\[
G(F) = \frac{g(F_0 + F) + g(F_0 - F)}{2}
\]

(6)

\( b_g \) is obtained by minimizing

\[
err^2_g = \int_{0}^{50} (G(F) - (a_g + b_gF))^2 dF
\]

(7)

over \( a_g \) and \( b_g \).

2.2. Estimating Vibrato Sections

Vibrato (deliberate, periodic fluctuation of F0) is considered as important singing technique, and so is incorporated within our scheme. Our vibrato detection scheme imposes restrictions on vibrato parameters of rate (the number of vibrations per second) and extent (the amplitude of vibration from an average pitch on the vibrato section). Restrictions of rate and extent are based on previous research [2, 10], with the rate range set at 5–8 Hz and extent range at 30–150 cent.

The basic idea is to detect vibrato by using short-term Fourier transform (STFT). In our current implementation, an STFT with a 32-sample (320 msec) Hanning window is calculated by using the Fast Fourier Transform (FFT). STFT is applied to \( \Delta F_{F0}(t) \), i.e. the first order finite differential of \( F_{F0}(t) \). The power spectrum \( X(f, t) \) can be expected to have a sharp peak where \( f \) corresponds to the vibrato rate. This is expressed by the power \( \Psi_v(t) \) and the sharpness \( S_v(t) \) defined as:

\[
\Psi_v(t) = \int_{F_L}^{F_H} \hat{X}(f, t) df
\]

(8)

\[
S_v(t) = \int_{F_L}^{F_H} \left| \frac{\partial \hat{X}(f, t)}{\partial f} \right| df
\]

(9)

\( 1 \) We avoid unnatural smoothing by ignoring silence sections and leaps of F0 transitions wider than a 300 cent threshold.

\[
\text{rate} = \frac{1}{N} \sum_{n=1}^{N} R_n
\]

(10)

\[
\text{extent} = \frac{1}{2N} \sum_{n=1}^{N} E_n
\]

(11)

\[
A \text{ section is judged as a vibrato section when it has high values of } P_v(t), \text{ and } \Delta F_{F0}(t) \text{ crosses its average value more than 5 times. The vibrato rate and extent are obtained by}
\]

(12)

(13)
Table 1: Dataset for classification experiment.

<table>
<thead>
<tr>
<th>name</th>
<th>class</th>
<th>language</th>
<th>gender</th>
<th>the number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>E004</td>
<td>good</td>
<td>English</td>
<td>female</td>
<td>50</td>
</tr>
<tr>
<td>E008</td>
<td>good</td>
<td>English</td>
<td>female</td>
<td>50</td>
</tr>
<tr>
<td>E017</td>
<td>good</td>
<td>English</td>
<td>male</td>
<td>50</td>
</tr>
<tr>
<td>E021</td>
<td>good</td>
<td>English</td>
<td>male</td>
<td>50</td>
</tr>
<tr>
<td>J002</td>
<td>good</td>
<td>Japanese</td>
<td>female</td>
<td>50</td>
</tr>
<tr>
<td>J054</td>
<td>good</td>
<td>Japanese</td>
<td>male</td>
<td>50</td>
</tr>
<tr>
<td>E001</td>
<td>poor</td>
<td>English</td>
<td>female</td>
<td>50</td>
</tr>
<tr>
<td>E002</td>
<td>poor</td>
<td>English</td>
<td>female</td>
<td>50</td>
</tr>
<tr>
<td>E013</td>
<td>poor</td>
<td>English</td>
<td>male</td>
<td>50</td>
</tr>
<tr>
<td>E023</td>
<td>poor</td>
<td>English</td>
<td>male</td>
<td>50</td>
</tr>
<tr>
<td>J014</td>
<td>poor</td>
<td>Japanese</td>
<td>female</td>
<td>50</td>
</tr>
<tr>
<td>J052</td>
<td>poor</td>
<td>Japanese</td>
<td>male</td>
<td>50</td>
</tr>
</tbody>
</table>

3.3. Results and Discussion

Two experiments with different evaluation criteria were conducted over three dataset settings (male, female, and male-female). In each case, the feature value space was classified using Support Vector Machine (SVM) as the classifier. The two evaluation criteria are 10-fold cross-validation and leave-one-out cross-validation. The 10-fold cross-validation method uses 9/10 of the samples as the training set and 1/10 as the test set in each trial. The leave-one-out cross-validation, in our context, evaluates one sample as test data, and uses the rest as training data (excluding those with the same singer/melody as the test data).

3.3. Results and Discussion

Table 2 and Table 3 show the classification rates, precision rates and recall rates of each class for the 10-fold and the leave-one-out cross-validation respectively. The classification rate (C), precision rate (P) and the recall rate (R) are defined as follows, where i denotes the class of good (i = good) or poor (i = poor).

\[
P_i = \frac{\text{samples correctly classified as class}_{i}}{\text{samples classified as class}_{i}} \times 100
\]

\[
R_i = \frac{\text{samples correctly classified as class}_{i}}{\text{samples in class}_{i}} \times 100
\]

\[
C = \frac{\text{total number of samples}}{\text{total number of samples}} \times 100 = \frac{R_{\text{good}} + R_{\text{poor}}}{2}
\]

The overall high values of P_{good} and R_{poor} suggest that the classification is stringent towards judging as “good” (if judged as good, then it is likely to be “real good”).

Figure 5 shows the classification rate for singers from the male-female dataset. The results show that the classification of good samples has a relatively high number of errors. One reason

3.2. Experimental Setting

The song samples are taken from the AIST Humming Database (AIST-HDB) [11]. The AIST-HDB contains singing voices of 100 subjects who sung melodies of 100 excerpts from 50 songs in the RWC Music Database (Popular Music [12] and Music Genre [13]).

Table 1 shows the singing voice dataset used in the experiment. Our automatic classification experiment used 600 samples by 12 singers who sung 50 excerpts from either 25 Japanese or English songs, after listening to each excerpt five times. The 12 singer subjects (ID’s in the “name” column) were selected by the criteria of receiving a consistently high rating (“good”) or low rating (“poor”) in the rating experiment by human subjects, and all their samples were marked good/poor accordingly (given in the “class” column in the table).
Table 2: Results of the 10-fold criteria.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C</th>
<th>P_{good}</th>
<th>R_{good}</th>
<th>P_{poor}</th>
<th>R_{poor}</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>87.7%</td>
<td>95.2%</td>
<td>79.3%</td>
<td>82.3%</td>
<td>96.0%</td>
</tr>
<tr>
<td>female</td>
<td>80.3%</td>
<td>91.7%</td>
<td>66.7%</td>
<td>73.8%</td>
<td>94.0%</td>
</tr>
<tr>
<td>male-female</td>
<td>83.3%</td>
<td>90.3%</td>
<td>74.7%</td>
<td>78.4%</td>
<td>92.0%</td>
</tr>
</tbody>
</table>

Table 3: Results of the leave-one-out criteria.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C</th>
<th>P_{good}</th>
<th>R_{good}</th>
<th>P_{poor}</th>
<th>R_{poor}</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>87.7%</td>
<td>93.8%</td>
<td>80.7%</td>
<td>70.8%</td>
<td>94.7%</td>
</tr>
<tr>
<td>female</td>
<td>71.7%</td>
<td>74.8%</td>
<td>65.3%</td>
<td>58.0%</td>
<td>78.0%</td>
</tr>
<tr>
<td>male-female</td>
<td>83.5%</td>
<td>87.6%</td>
<td>78.0%</td>
<td>70.3%</td>
<td>89.0%</td>
</tr>
</tbody>
</table>

is that even a good singer may occasionally fail to keep good pitch intervals, as they were singing out from memory. Such cases include when the overall pitch undergoes a gradual drift (resulting in bad ratings in 1 and 2), which to the human ear does not sound so distorted.

Classification errors also arise from the vibrato features. A “good” sample can be misclassified as poor when there is no vibrato, or when the values surpass the range restriction of vibrato parameters. On the other hand, poor singing which cannot keep a stable F0 can be mistakenly judged as a vibrato, especially resulting from relatively high values of vibrato.

4. Conclusion

This paper proposed two acoustical features, pitch interval accuracy and vibrato, which are effective for evaluating singing skills without score information. In addition to these features, we have also investigated other features, such as the slope of long-term average spectrum (LTAS), the power within a frequency range of singer’s formant, the variance of 16-order cepstral coefficients, the ratio between the power of harmonic components and others, the average of spectral centroid or spectral rolloff, and the standard deviation of F0 or power. The performance of their combinations, however, has not surpassed the performance of the two features presented in this paper. In the future, we plan to investigate other features relevant to different vocal aspects such as rhythm and vocal quality.

5. Acknowledgements

Authors would like to thank Mr. Hirokazu Kameoka (the University of Tokyo) for his valuable discussions and Dr. Elias Pampalk (CREST/AIST) for proofreading an earlier version.

6. References


