

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

Tomoyasu Nakano

Masataka Goto

Yuzuru Hiraga

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

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

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 (F_0) 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 \boxed{n} . 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 F_0 (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_{cent} of frequency f_{Hz} given as follows (middle C corresponds to 4800 cent).

$$f_{\text{cent}} = 1200 \log_2 \frac{f_{\text{Hz}}}{440 \times 2^{\frac{3}{12} - 5}} \quad (1)$$

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 ω_i is a weight factor (currently set to 1), and σ_i ($= 16$ cent) is the standard deviation of the Gaussian distribution.

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

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 $F_{F_0}(t)$ denotes the F_0 and $P_{F_0}(t)$ denotes the F_0 possibility at time t , which are estimated per 10 msec using the method of Goto et al. [9].

$$P_g(F, t) = \int_{t-T_g}^t p(F_{F_0}(\tau); F) P_{F_0}(\tau) d\tau \quad (3)$$

If all the F_0 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 $P_g(F, t)$ distribution thus indicates the degree of deviation of the singer from any semitone grid.

Figure 1 shows the calculation process and example results of $P_g(F, t)$. 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 frequency¹, 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]. $P_g(F, t)$ is calculated with a 200-sample (2sec) 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 $P_g(F, t)$ values shown to the right.

Figure 2 shows examples of $P_g(F, t)$ and its long-term average $g(F)$. When the singing is “good”, then $P_g(F, t)$, 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]:

$$M = \int_{F_g-50}^{F_g+50} (F_g - F)^2 g(F) dF \quad (4)$$

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

$$F_g = \operatorname{argmax}_F g(F) \quad (5)$$

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

$$G(F) = \frac{g(F_g + F) + g(F_g - F)}{2} \quad (6)$$

b_g is obtained by minimizing

$$\operatorname{err}_g^2 = \int_0^{50} (G(F) - (a_g + b_g F))^2 dF \quad (7)$$

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:

$$\Psi_v(t) = \int_{F_L}^{F_H} \hat{X}(f, t) df \quad (8)$$

$$S_v(t) = \int_{F_L}^{F_H} \left| \frac{\partial \hat{X}(f, t)}{\partial f} \right| df \quad (9)$$

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

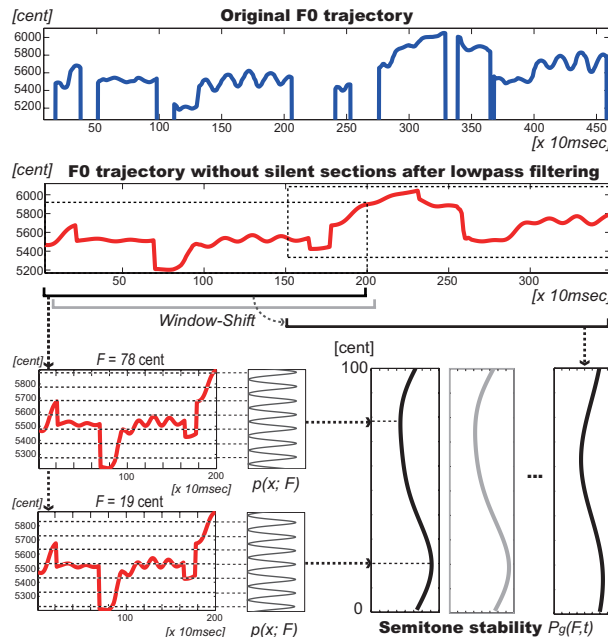


Figure 1: Overview of calculation method of $P_g(F, t)$.

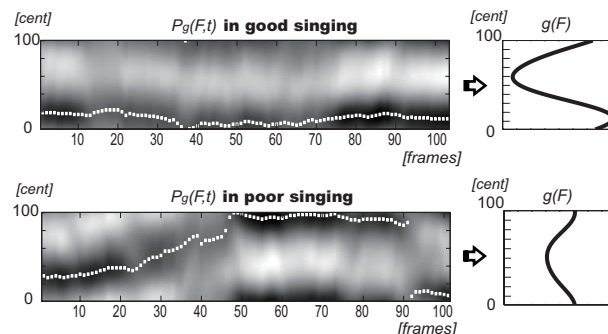


Figure 2: Examples of $P_g(F, t)$ and its long-term average $g(F)$ in good/poor singing.

where (F_L, F_H) is the range of vibrato rate, and $\hat{X}(f, t)$ is $X(f, t)$ normalized over f :

$$\hat{X}(f, t) = \frac{X(f, t)}{\int X(f, t) df} \quad (10)$$

Using these, the vibrato likeliness $P_v(t)$ is defined as

$$P_v(t) = S_v(t) \Psi_v(t) \quad (11)$$

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

$$\frac{1}{\text{rate}} = \frac{1}{N} \cdot \sum_{n=1}^N R_n \quad (12)$$

$$\text{extent} = \frac{1}{2N} \cdot \sum_{n=1}^N E_n \quad (13)$$

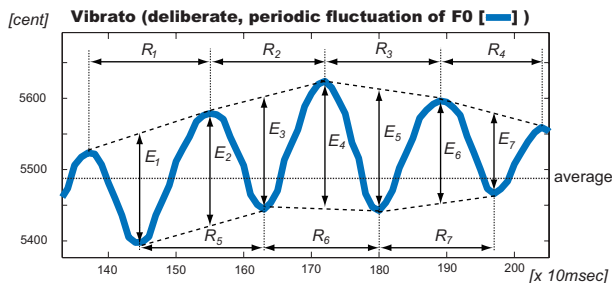


Figure 3: Extraction parameters for vibrato detection.

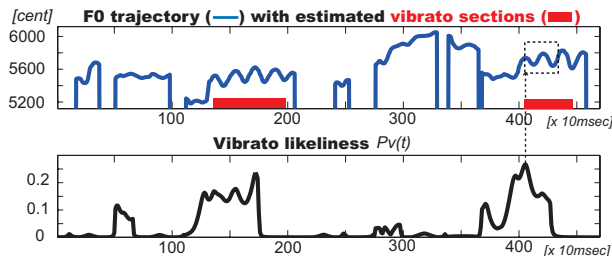


Figure 4: An example of vibrato detection.

where R_n [sec] and E_n [cent] are illustrated in Figure 3. The total length of the vibrato section is used as feature [3].

The following two functions P_{v1} and P_{v2} are also used as features [4] and [5], since they take high values in the presence of vibratos (T_v is the length of the analyzed signal).

$$P_{v1} = \max_{0 \leq t \leq T_v} (P_v(t)) \quad (14)$$

$$P_{v2} = \frac{1}{T_v} \int_0^{T_v} P_v(t) dt \quad (15)$$

3. Classification Experiment

The proposed features [1]–[5] have been tested in 2-class classification (*good/poor*) experiments, using the results of a previous rating experiment by human subjects [8].

3.1. Dataset

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 1: Dataset for classification experiment.

name	class	language	gender	the number of samples
E004	good	English	female	50
E008	good	English	female	50
E017	good	English	male	50
E021	good	English	male	50
J002	good	Japanese	female	50
J054	good	Japanese	male	50
E001	poor	English	female	50
E002	poor	English	female	50
E013	poor	English	male	50
E023	poor	English	male	50
J014	poor	Japanese	female	50
J052	poor	Japanese	male	50

3.2. Experimental Setting

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_i) and the recall rate (R_i) are defined as follows, where i denotes the class of good ($i = good$) or poor ($i = poor$).

$$P_i = \frac{\text{samples correctly classified as class}_i}{\text{samples classified as class}_i} \times 100 \quad (16)$$

$$R_i = \frac{\text{samples correctly classified as class}_i}{\text{samples in class}_i} \times 100 \quad (17)$$

$$C = \frac{\text{samples correctly classified}}{\text{total number of samples}} \times 100 = \frac{R_{good} + R_{poor}}{2} \quad (18)$$

The results for male and male-female datasets are similar in both criteria, while there is a significant drop of ratings in the leave-one-out criteria for the female dataset. The female dataset also has a slightly lower rating in the 10-fold criteria as well. The reason for this drop is yet unclear, although the overall agreement in the male-female dataset suggests that the proposed features are both effective, and also robust against individual difference of singer and/or melody.

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 have a relatively high number of errors. One reason

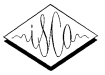


Table 2: Results of the 10-fold criteria.

Dataset	C	P_{good}	R_{good}	P_{poor}	R_{poor}
male	87.7%	95.2%	79.3%	82.3%	96.0%
female	80.3%	91.7%	66.7%	73.8%	94.0%
male-female	83.3%	90.3%	74.7%	78.4%	92.0%

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

Dataset	C	P_{good}	R_{good}	P_{poor}	R_{poor}
male	87.7%	93.8%	80.7%	70.8%	94.7%
female	71.7%	74.8%	65.3%	58.0%	78.0%
male-female	83.5%	87.6%	78.0%	70.3%	89.0%

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 2]), which to the human ear does not sound so distorted.

Classification errors also arise from the vibrato features [3–5]. 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 [4 5].

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

[1] Saitou, T., Unoki, M. and Akagi, M., “Development of an F0 Control Model Based on F0 Dynamic Characteristics for Singing-voice Synthesis”, *Speech Communication*, 46: 405–417, 2005.

[2] Sundberg, J., *The Science of the Singing Voice*, Northern Illinois Univ Press, 226p., 1987.

[3] Kawahara, H. and Katayose, H., “Scat Generation Research Program Based on STRAIGHT, a High-quality Speech Anal-

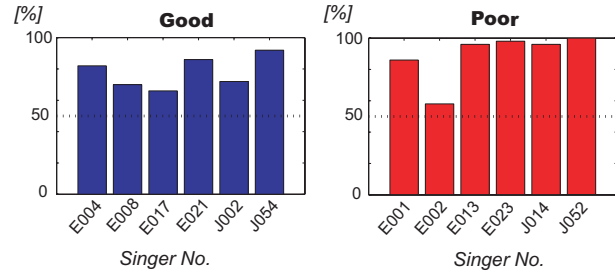


Figure 5: Result of the classification experiment (by singer) using the leave-one-out criteria and the male-female dataset.

ysis, Modification and Synthesis System”, *IPSP Journal*, 43(2):208–218, 2002. (in Japanese)

[4] Ohishi, Y., Goto, M., Ito, K. and Takeda, K., “Discrimination between Singing and Speaking Voices”, in *Proc. 9th European Conference on Speech Communication and Technology (Interspeech2005)*, 1141–1144, 2005.

[5] Omori, K., Kacker, A., Carroll, L.M., Riley, W.D. and Blaugrund, S.M., “Singing Power Ratio: Quantitative Evaluation of Singing Voice Quality”, *Journal of Voice*, 10(3):228–235, 1996.

[6] Brown, W. S. Jr., Rothman, H. B. and Sapienza, C.M., “Perceptual and Acoustic Study of Professionally Trained Versus Untrained Voices”, *Journal of Voice*, 14(3):301–309, 2000.

[7] Watts, C., Barnes-Burroughs, K., Estis, J. and Blanton, D., “The Singing Power Ratio as an Objective Measure of Singing Voice Quality in Untrained Talented and Nontalented Singers”, *Journal of Voice*, 20(1):82–88, 2006.

[8] Nakano, T., Goto, M. and Hiraga, Y., “Subjective Evaluation of Common Singing Skills Using the Rank Ordering Method”, in *Proc. 9th International Conference of Music Perception and Cognition (ICMPC2006)*, 2006. (accepted)

[9] Goto, M., Itou, K. and Hayamizu, S., “A Real-time Filled Pause Detection System for Spontaneous Speech Recognition”, in *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech '99)*, 227–230, 1999.

[10] Seashore, C.E., “A Musical Ornament, the Vibrato”, Chapter 4, in *Psychology of Music*, McGraw-Hill Book Company, pp.33–52, 1938.

[11] Goto, M. and Nishimura, T., “AIST Humming Database: Music Database for Singing Research”, *The Special Interest Group Notes of IPSJ (MUS)*, 2005(82):7–12, 2005. (in Japanese)

[12] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R., “RWC Music Database: Popular, Classical, and Jazz Music Databases”, in *Proc. 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, 287–288, 2002.

[13] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R., “RWC Music Database: Music Genre Database and Musical Instrument Sound Database”, in *Proc. 4th International Conference on Music Information Retrieval (ISMIR 2003)*, 229–230, 2003.