9th International Conference on Music Perception and Cognition

Alma Mater Studiorum University of Bologna, August 22-26 2006

Subjective evaluation of common singing skills using the rank ordering method

Tomoyasu Nakano
Graduate School of Library,
Information and Media Studies,
University of Tsukuba
Tsukuba, Ibaraki 305-8550, Japan
nakano@slis.tsukuba.ac.jp

Masataka Goto
National Institute of
Advanced Industrial Science
and Technology (AIST)
Tsukuba, Ibaraki 305-8568, Japan
m.goto@aist.go.jp

Yuzuru Hiraga
Graduate School of Library,
Information and Media Studies,
University of Tsukuba
Tsukuba, Ibaraki 305-8550, Japan
hiraga@slis.tsukuba.ac.jp

ABSTRACT

This paper presents the results of two experiments on singing skill evaluation, where human subjects judge the subjective quality of previously unheard melodies. The aim of this study is to explore the criteria that human subjects use in judging singing skill and the stability of their judgments, as a basis for developing an automatic singing skill evaluation scheme.

The experiments use the rank ordering method, where the subjects ordered a group of given stimuli according to their preferred rankings. Experiment 1 uses real, a capella singing as the stimuli, while experiment 2 uses the fundamental frequency (F0) sequence extracted from the singing. In experiment 1, 88.9% of the correlation between the subjects' evaluations was significant at the 5% level. Results of experiment 2 show that the F0 sequence is significant in only certain cases, so that the judgment and its stability in experiment 1 should be attributed to other factors of real singing.

Keywords

singing skill, subjective evaluation, rank ordering method

BACKGROUND

Automatic evaluation of singing skills is a promising research topic with various applications in scope. Previous research on singing evaluation has focused on trained, professional singers (mostly in classic music), using various approaches from physiology, anatomy, acoustics, and psychology – with the aim of presenting objective, quantitative measures of singing quality. Such works have reported that the singing voices have singer's formant [1] and the specific characteristics of fundamental frequency (*F0*) [2]. In particular, the singer's formant characterizes singing quality as *ringing* [3].

Our interest is directed more towards ordinary, common person's singing, understanding how they mutually evaluate their quality, and to incorporate such findings in an automatic evaluation scheme.

AIMS

The aim of this study is to explore the criteria that human subjects use in judging singing skill, and identify whether such judgments are stable and in mutual agreement among different subjects. This will serve as a preliminary basis for our goal of developing an automatic singing skill evaluation scheme.

Two experiments were carried out. Experiment 1 is intended to verify the stability of human judgment, using *a capella* singing sequences (*solo singing*) as the stimuli. Experiment 2 uses the F0 sequences *(F0 singing*) extracted from solo singing, and is intended to identify their contribution in the judgment. In both experiments, the melodies were previously unheard by the subjects.

METHOD AND EXPERIMENTS

The standard method of subjective evaluation by giving grade scores to each tested stimuli [4] is inappropriate for our case of singing evaluation, where the subtleties of subjects' judgments may be obscured by differences in musical experience. So instead, we used a rank ordering method, where the subjects were asked to order a group of stimuli according to their preferred rankings.

The singing samples are digital recordings of 16bit/16kHz/monaural. In order to suppress the variance between the samples, all the samples were set at the same volume and were presented through a headphone.

Interface for Subjective Evaluation

Figure 1 shows the interface screen used in the experiments. The speaker icons indicate 10 stimuli (A, B, ..., J), which can be double-clicked to play the sound, and can be moved around by drag-and-drop using the mouse. The subjects are instructed to align the icons horizontally according to their order of judgment, ranging from *poor* (left-hand side) to *good* (right-hand side). The vertical positioning is insignificant for the experiments.

The left figure shows an initial setting (random order), and the right figure shows an example result, with H judged as the best and B as the poorest. At the end of the experiment, the subjects are also instructed to insert two lines (1. and 2. in the right figure) classifying the samples into "good" (H, I), "poor" (B, F, D) and "intermediate" (E, J, A, G, C).

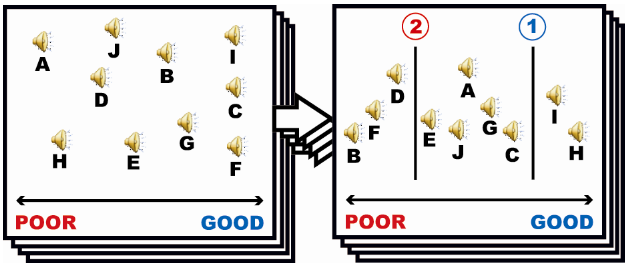

Figure 1. Example subjective evaluation session using the interface screen..

The Measurement of Rank Correlation

The results are analyzed using the Spearman's rank correlation ρ [5] defined as follows:

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

where N is the number of stimuli (= 10 in the experiments), and a_i, b_i are the i-th component (rank value) of the rank vectors **a** and **b**. The value of ρ ranges from 1 (**a** = **b**) to −1 (**a**, **b** are reverse order). The correlation of **a**, **b** is significant at the 1% level for $\rho \geq 0.7333$ and 5% level for $\rho \geq 0.5636$ [5].

Experiment 1

This experiment uses solo singing as the stimuli. The subjects were presented with four groups of singing, each group with the same melody sung by 10 singers. The task is to order each group using the interface explained above. The subjects were free to listen to the melodies as many times as they want to.

The subjects were also asked to give introspective description of their judgments.

Subjects

22 subjects (University students, ages 19 to 29) participated in the experiment. 16 had experience with musical instruments, and 2 had experience with vocal music (popular or chorus). 4 stated to possess absolute pitch. The subjects were divided into two sets (A, B, each with 11 subjects), each set presented with the same stimuli set.

Stimuli

The samples of stimuli were taken from the *RWC Music Database: Popular Music* (RWC-MDB-P-2001) [7] and the *AIST Humming Database* (AIST-HDB) [6]. The AIST-HDB contains singing voices of 100 subjects, each singing the melodies of two excerpts from the *chorus* and *verse A* sections of 50 songs (100 samples) in the RWC Music Database (Popular Music [7] and Music Genre [8]) Table 1 shows the two stimuli sets A and B. Each set has 4 different melodies, sung by 10 individuals of the same gender (1 from RWC-MDB-P and 9 from AIST-HDB) presented as a group on the interface screen. The language of the lyrics is either Japanese or English.

Experiment 2

Experiment 2 follows the same procedure as experiment 1, except that the stimuli are replaced with F0 singing, extracted from the solo singing used in experiment 1 (see below). The subjects were further instructed to ignore any noise cased by the F0 extraction process.

Subjects

20 subjects (University students, ages 19 to 35) participated in the experiment. None of them participated in experiment 1. 17 had experience with musical instruments, and 6 had experience with vocal music (popular or chorus). 6 stated

to possess absolute pitch. The subjects were divided into two sets (A, B, each with 10 subjects), each set presented with the same stimuli set.

Stimuli

The stimuli used in this experiment are F0 sequences extracted from the samples used in experiment 1, removing all other vocal features. F0 is estimated per 10 msec using the method of Goto et al. [9], and is resynthesized as a sinusoidal wave with its amplitude preserving the power of the most predominant harmonic structure of the original. The resulting F0 sequence a natural impression comparable to the original.

Table 1. 80 stimuli (by 40 singers).

set	music No.	excerpted section	lyrics language	gender	the number of singers
A	27	verse A	Japanese	male	10
	28	verse A	Japanese	female	10
	90	verse A	English	male	10
	97	chorus	English	female	10
B	27	chorus	Japanese	male	10
	28	chorus	Japanese	female	10
	90	chorus	English	male	10
	97	verse A	English	female	10

Note: music No. are from RWC-MDB-P-2001

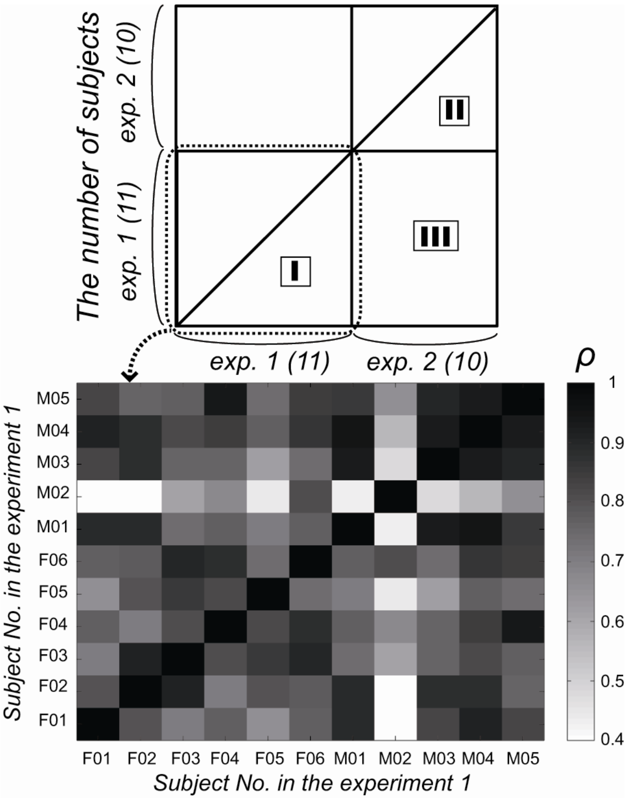

Figure 2. Graphical scheme of the ρ-matrix (above) and example gradation display for the *A—English—female* case.

RESULTS

The ranking correlation (1) is calculated for all pairs of subject rankings, giving the ρ-matrix shown schematically in Figure 2. In the top figure, $\boxed{\text{I}}$ corresponds to pairings for rankings in experiment 1 (55 = 11 x (11 - 1)/2 pairs), $\boxed{\text{II}}$ to pairings in experiment 2 (45 = 10 x (10 - 1)/2 pairs), and $\boxed{\text{III}}$ to cross-pairings for rankings in experiments 1 and 2 (110 = 11 x 10 pairs). The bottom figure shows an example gradation display of the ρ-matrix for the "$\boxed{\text{I}}$ *-A-English-female*" case, corresponding to set A, English lyrics, female singer group. The gradation is darker for higher ρ values.

Tables 2,3 show the results for $\boxed{\text{I}}$ (ρ values of solo singing). Table 2 shows the percentage of significant pairs, and Table 3 shows the statistical values of ρ for each of the groups. The results show the stability of subject judgments for solo singing.

Each singing sample was further labeled as good, poor or otherwise (which is to be used for developing the automatic evaluation scheme), using the following criteria:

good many subjects evaluated the sample as good and no subject as poor,

poor many subjects evaluated the sample as poor and no subject as good,

otherwise neither of the above.

Table 4 shows the results of labeling.

Table 2. Percentage of significant pairs (solo singing).

set	lyrics language	gender	p<.01	p<.05
A	Japanese	male	96.4% (53)	100.0% (55)
	Japanese	female	74.6% (41)	90.9% (50)
	English	male	61.8% (34)	89.1% (49)
	English	female	41.8% (23)	80.0% (44)
	overall (220)		68.6% (151)	90.0% (198)
B	Japanese	male	45.5% (25)	72.7% (40)
	Japanese	female	72.7% (40)	98.2% (54)
	English	male	52.7% (29)	89.1% (49)
	English	female	74.6% (41)	90.9% (50)
	overall (220)		61.4% (135)	87.8% (193)
overall (440)			65.0% (260)	88.9% (391)

Table 3. Statistics of ρ (solo singing).

set	lyrics language	gender	mean (SD)	min / max
A	Japanese	male	0.87 (0.07)	0.71 / 0.99

	Japanese	female	0.77 (0.14)	0.38 / 0.95
	English	male	0.75 (0.14)	0.28 / 0.96
	English	female	0.69 (0.14)	0.42 / 0.98
B	Japanese	male	0.64 (0.22)	0.03 / 0.98
	Japanese	female	0.81 (0.13)	0.39 / 0.99
	English	male	0.73 (0.14)	0.36 / 0.98
	English	female	0.76 (0.14)	0.36 / 0.96

Table 4. Results of labeling (good/poor).

set	lyrics language	gender	good	poor	otherwise
A	Japanese	male	3/10	2/10	5/10
	Japanese	female	3/10	3/10	4/10
	English	male	4/10	2/10	4/10
	English	female	3/10	2/10	5/10
B	Japanese	male	1/10	3/10	7/10
	Japanese	female	3/10	3/10	4/10
	English	male	2/10	2/10	6/10
	English	female	3/10	4/10	3/10

Table 5. Percentage of significant pairs (F0 singing).

set	lyrics language	gender	p<.01	p<.05
A	Japanese	male	44.4% (20)	77.8% (35)
	Japanese	female	55.6% (25)	71.1% (32)
	English	male	15.6% (7)	37.8% (17)
	English	female	17.8% (8)	35.6% (16)
	overall (180)		33.3% (60)	55.6% (100)
B	Japanese	male	2.2% (1)	13.3% (6)
	Japanese	female	22.2% (10)	44.4% (20)
	English	male	15.6% (7)	46.7% (21)
	English	female	40.0% (18)	62.2% (28)
	overall (180)		20.0% (36)	41.7% (75)
overall (360)			26.7% (96)	48.6% (175)

Table 6. Statistics of ρ (F0 singing).

set	lyrics language	gender	mean (SD)	min / max
A	Japanese	male	0.68 (0.17)	0.22 / 0.94
	Japanese	female	0.69 (0.18)	0.26 / 0.94
	English	male	0.44 (0.27)	-0.24 / 0.89
	English	female	0.32 (0.39)	-0.87 / 0.88
B	Japanese	male	0.27 (0.27)	-0.33 / 0.79
	Japanese	female	0.52 (0.23)	-0.03 / 0.89
	English	male	0.45 (0.29)	-0.21 / 0.87
	English	female	0.64 (0.16)	0.26 / 0.94

Table 7. Percentage of significant pairs (solo—F0 singing).

set	lyrics language	gender	p<.01	p<.05
A	Japanese	male	54.5% (60)	82.7% (91)
	Japanese	female	40.0% (44)	78.2% (86)
	English	male	25.5% (28)	56.4% (62)
	English	female	20.9% (23)	55.5% (61)
	overall (220)		35.2% (155)	68.2% (300)
B	Japanese	male	12.7% (14)	27.3% (30)
	Japanese	female	29.1% (32)	60.0% (66)
	English	male	21.8% (24)	44.5% (49)
	English	female	41.8% (46)	78.2% (86)
	overall (220)		26.4% (116)	52.5% (231)
overall (440)			30.1% (271)	60.3% (531)

Table 8. Statistics of ρ (solo—F0 singing).

set	lyrics language	gender	mean (SD)	min / max
A	Japanese	male	0.72 (0.17)	0.24 / 0.98
	Japanese	female	0.66 (0.18)	0.13 / 0.92
	English	male	0.56 (0.24)	-0.21 / 0.99
	English	female	0.48 (0.32)	-0.44 / 0.90
B	Japanese	male	0.42 (0.26)	-0.27 / 0.90
	Japanese	female	0.61 (0.19)	0.08 / 0.92
	English	male	0.50 (0.24)	-0.10 / 0.98
	English	female	0.66 (0.17)	0.14 / 0.95

Tables 5,6 corresponds to the results of Tables 2,3 for II (F0 singing), and Tables 7,8 for III (solo—F0 singing cross correlation). The results for II show the stability of subject judgments for F0 singing, while the results for III show the correlation between judgments for solo and F0 singing, indicating the amount of contribution of the F0 factor. Figure 3 shows the bar graph indicating that the results of Tables 2, 5, 7.

The criteria that human subjects use in judging singing skill can also be looked into from the introspective comments. Example features mentioned in the comments for experiment 1 include:

- tonal stability
- rhythmical stability
- pronunciation quality
- singing technique (*e.g. vibrato, keeping a stable F0*)
- vocal expression and quality
- good/poor can be classified from a short sequence (3—5 seconds)
- personal preference

Likewise for experiment 2:

- tonal stability
- rhythmical stability
- singing technique (*e.g. vibrato, keeping a stable F0*)
- vocal expression and quality

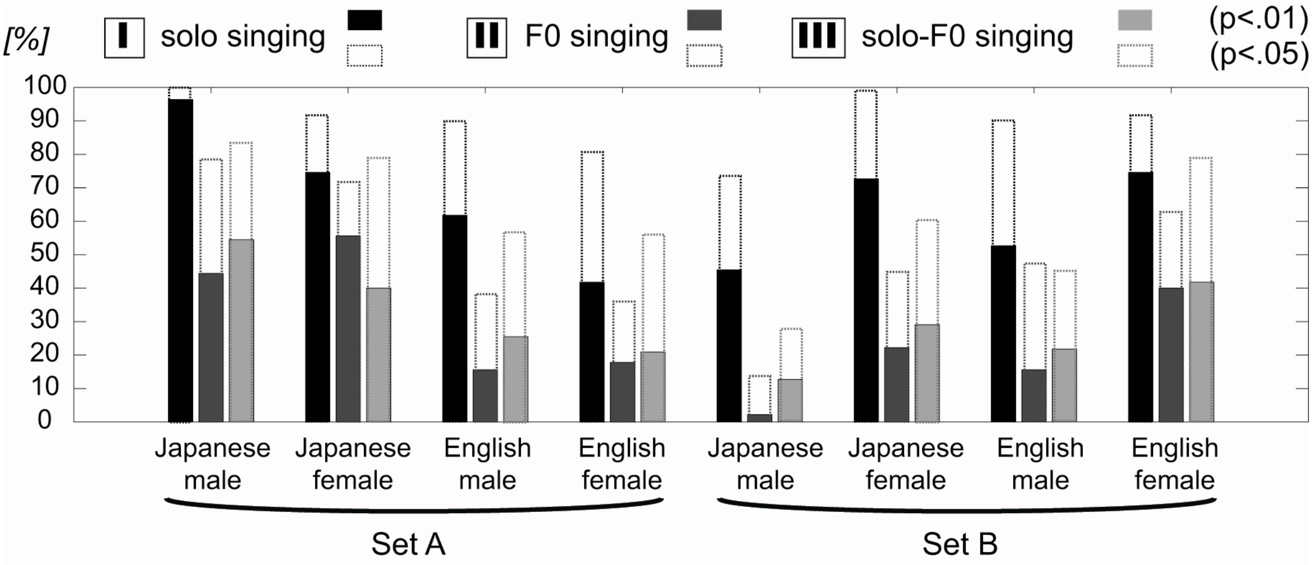

Figure 3. Percentage of significant pairs.

DISCUSSION

The results of I show that 391 pairs (88.9%) of subject rankings were significant at the 5% level, and 260 pairs (65.0%) were significant at the 1% level. This suggests that the rankings are generally stable and in mutual agreement, meaning that they are based more on common, objective features, contrary to the comments mentioning that evaluation is a matter of personal preference. The ρ values in Tables 3, 6, 8 all have positive (and in many cases, high) mean values, also indicating that the general tendency of the rankings are stable. Furthermore, in the good/poor classification, none of the samples were completely divided between good and poor ratings.

Being such, the results of the labeling (good/poor) can be taken as a sufficiently reliable basis to be utilized in developing an automatic evaluation scheme. This is further supported by the fact that many comments refer to objective (or at least, objectively taken) features such as tonal stability as judgment criteria, and that only a short sequence (3—5 sec.) is sufficient for judging good/poor. These points give practical support for the realizability of such a scheme.

The results of II show that the subjects' rankings of F0 singing are stable in some cases (*e.g. A—Japanese—male, A—Japanese—female*, and *B—English—female*) but not so in others. High correlation rates are obtained when the melodies consist of relatively long notes, which require higher singing skills. But together with the relatively low overall values of the results of III, it can be said that F0 alone is not decisive for judging singing skills, and other acoustic and musical features are incorporated in achieving the high correlation rates in the results of I. One interesting point is that some comments for experiment 2 mentioned "vocal expression or quality", indicating that such features can (at least in a subjective sense) be recognized even with information of F0 alone.

ISBN 88-7395-155-4 © 2006 ICMPC

CONCLUSION

The results show that under the control of lyrics language, singers' gender, and melody type (verse/chorus), the rankings given by the subjects are generally stable, indicating that they depend more on common, objective features rather than reflecting subjective preference. This makes the results reliable enough to be used as a referendum for developing automatic singing evaluation schemes.

Further experiments will be conducted in various other settings to explore singing skills in more detail. Work on identifying the key acoustic properties that underlie human judgments is also in progress.

ACKNOWLEDGMENTS

This paper utilized the RWC Music Database (Popular Music) and AIST Humming Database.

REFERENCES

[1] Sundberg, J. (1987). *The Science of the Singing Voice*. Illinois: the Northern Illinois University Press.

[2] Saitou, T., Unoki, M. & Akagi, M. (2005). Developmentof an F0 Control Model Based on F0 Dynamic Characteristics for Singing-voice Synthesis. *Speech Communication, 46,* 405-417.

[3] Omori, K., Kacker, A., Carroll, L. M., Riley, W. D. & Blaugrund, S. M. (1996). Singing Power Ratio: Quantiative Evaluation of Singing Voice Quality. *Journal of Voice, 10 (3),* 228-235.

[4] Franco, H., Leonardo, N., Digalakis, V. & Ronen, O. (2000). Combination of Machine Scores for Automatic Grading of Pronunciation Quality. *Speech Communication, 30,* 121-130.

[5] Kendall, M. & Gibbons, J. D. (1990). *Rank Correlation Methods*. New York: Oxford University Press.

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

[7] Goto, M., Hashiguchi, H., Nishimura, T. & Oka, R. (2002). RWC Music Database: Popular, Classical, and Jazz Music Databases. in *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR2002),* 287-288.

[8] Goto, M., Hashiguchi, H., Nishimura, T. & Oka, R. (2003). RWC Music Database: Music Genre Database and Musical Instrument Sound Database. in *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR2003),* 229-230.

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