LETTER

Automatic Singing Performance Evaluation for Untrained Singers*

Chuan CAO[†a)], Ming LI[†], Xiao WU[†], *Nonmembers*, Hongbin SUO[†], *Student Member*, Jian LIU[†], *and* Yonghong YAN[†], *Nonmembers*

SUMMARY In this letter, we present an automatic approach of objective singing performance evaluation for untrained singers by relating acoustic measurements to perceptual ratings of singing voice quality. Several acoustic parameters and their combination features are investigated to find objective correspondences of the perceptual evaluation criteria. Experimental results show relative strong correlation between perceptual ratings and the combined features and the reliability of the proposed evaluation system is tested to be comparable to human judges.

key words: *automatic/objective evaluation, singing performance assessment, feature combination*

1. Introduction

Singing performance evaluation was once thought to be inherently subjective and listener-dependent and it has received little research attention for a long time. Even in the perceptual domain, defined standards for singing performance evaluation were fairly scarce. However, several recent investigations have made significant contributions to the study of the auditory-perceptual evaluation of solo singing performance. Wapnick & Ekholm found a set of perceptual evaluation terms whose validity was established through consensus among expert singing teachers and extensive review of the pedagogical literature [1]. Oates *et al* [2] did further investigation on similar perceptual terms and developed an auditory-perceptual rating instrument for operatic singing voice. In our recent study [3], we investigated several perceptual criteria for untrained singers' singing performance evaluation which intuitively have strong physical and acoustic implications and critical evaluation criteria were found for further studies.

As to objective singing evaluation aspect, Bartholomew indicated some physical definitions of good voice-quality of male voice early in 1934 [4]. Later in 1998, Ekholm *et al* had a study of relating objective measurements to expert singing performance ratings in [5]. Much deeper in the signal processing field, Sundberg found and stated several acoustic features that related to vocal ugliness [6] and Nakano *et al* presented a 2-class (*good/poor*) automatic singing skill evaluation method for unknown-melody songs [7]. However, most acoustic measurements in previous works were obtained by semi-automated procedures and they did not draw a clear conclusion on how the acoustic measurements relate to the perceptual criteria. And Nakano's system could only handle good/poor classification problems and they did not explore the usefulness of target-song information such as melody and tempo.

In this letter, we present a study of the relationship between objective acoustic measurements and perceptual ratings of singing performance evaluation. Feature combination techniques are used in this study, aiming to find more relevant correspondences of subjective perception, by the method of integrating several relevant acoustic parameters into one combined feature. Finally, we present an automatic singing quality evaluation approach, which is able to automatically map acoustic measurements to singing performance ratings. Two main experiments on a singing-clips dataset are conducted to test the system performance. One aims to measure how much the significant perceptual criteria correlate with their corresponding acoustic cues, including combined features. The other is conducted to verify the reliability of the proposed automatic objective singing evaluation approach, compared to human expert judges. Such objective singing performance evaluation systems can be applied in many applications, such as singing pedagogy, voice training feedback systems and entertainment singing contest. More importantly, this kind of trials could help expand signal processing & Music Information Retrieval (MIR) techniques to the artistic scope and bridge the gap between vocal artists and scientists. Since different music styles have their own characteristics, we focus on the pop-style singing of untrained singers in this work. But we think similar frameworks can also be appropriately utilized to most of other singing styles.

2. Perceptual Evaluation Criteria

In our previous study [3], several perceptual evaluation criteria (selected from consensus proposed in [2], [5]) are found to have intuitively strong acoustic implications and also have critical impact on rating singing performance for untrained singers during short singing clips. Concise descriptions of these perceptual terms are:

Manuscript received November 4, 2008.
Manuscript revised March 11, 2009.

[†]The authors are with the ThinkIT Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, China.

[*]This work is partially supported by The National High Technology Research and Development Program of China (863 program, 2006AA010102, 2006AA01Z195), MOST (973 program, 2004CB318106), National Natural Science Foundation of China (10574140, 60535030).

a) E-mail: ccao@hccl.ioa.ac.cn

DOI: 10.1587/transinf.E92.D.1596

- Intonation accuracy: singing in tune with matching pitch
- Rhythm consistency: singing with appropriate speed consistent with the origin
- Timbre brightness: brilliance of tone, a sensation of brightness, ring and warmth
- Vocal clarity: a sensation of feeling the vocal vibrations of a clear, well-produced tone
- Overall performance: the overall evaluation integrating all perceptual terms

Intuitively, "appropriate vibrato" could also be an important aspect of singing evaluation. But after analyzing the experiment data we found that vibrato rarely appears in the corpus, which is a normal phenomenon for untrained singers' pop-style singing voice [3]. So we focus on the previous five criteria in this work.

3. Corresponding Acoustic Measurements

Generally, singing is mainly an acoustic activity. Human subjects are able to evaluate the singing quality only by listening to sound recordings. So there should be plenty of critical acoustic cues related to the perceptual assessment criteria described in Sect. 2.

3.1 Intonation Accuracy

In this study, the perception of "intonation accuracy" is intuitively related to pitch accuracy, which refers to the similarity between singing subject's pitch curve and that of the original singer of the target song. Among the state-of-art similarity measurement methods, we investigate Wu's frame-level Recursive Alignment (RA) algorithm [8] and a classic note-level Dynamic Time Warping (DTW) method. Wu's algorithm can be concisely described as: given two F0 (fundamental frequency) sequences, $A_{1\cdots m}$ and $B_{1\cdots n}$, their similarity can be iteratively calculated by:

$$\begin{aligned} C_{RA}(A_{1\cdots m}, B_{1\cdots n}) = C_{RA}(&A_{1\cdots k}, B_{1\cdots l}) \\ &+ C_{RA}(A_{k\cdots m}, B_{l\cdots n}) \end{aligned} \tag{1}$$

where k and l are selected by:

$$\begin{aligned} <k,l> = \arg\max_{k,l}\{&LS(A_{1\cdots k}, B_{1\cdots l}) \\ &+ LS(A_{k\cdots m}, B_{l\cdots n}) | k \in [1,m], l \in [1,n]\} \end{aligned} \tag{2}$$

where LS refers to the linear scaling cost described in [9], which could be simply understood as a warping Euclidean distance. The F0 estimation method used in this work is similar to our previous work in music melody extraction [10], which is based on the subharmonic summation framework and a harmonic structure tracking strategy.

A classic DTW alignment cost is also investigated. Our measurement is on the note-level, and the specific implementation can be expressed as: given two note sequences, $A_{1\cdots N_A}$ and $B_{1\cdots N_B}$, the DTW cost of A-align-to-B is measured by:

$$C_{DTW}(A_{1\cdots N_A}, B_{1\cdots N_B}) = \min\{D_{N_A,j} | j \in [1, N_B]\} \tag{3}$$

in which,

$$D_{i,j} = d(A_i, B_j) + \min(D_{i-1,j-1}, D_{i-1,j-2}, D_{i-2,j-1}) \tag{4}$$

where, $d(A_i, B_j)$ represents the distance between note A_i and note B_j, and $D_{i,j}$ means the minimum cumulative cost up to A_i and B_j.

The note sequences of the singing subjects are obtained through a energy based note segmentation method, which is detailedly described in [11]. Since F0 and note sequences of the original singer are hard to obtain, MIDI files of the target songs are used as references in this study.

3.2 Rhythm Consistency

We relate the perception "rhythm consistency" to the objective term "rhythm accuracy", which is a more common terminology used by MIR researchers. In our implementation, rhythm accuracy is measured by the note duration consistency between the singing subject and the target MIDI. Given note duration sequence $D_{A_i\cdots A_N}$, and its corresponding sequence $D_{B_i\cdots B_N}$, two measurements are investigated: one is the cosine distance,

$$M_{Cos} = \left(\sum_{i=1}^{N}(D_{A_i} \cdot D_{B_i})\right) \Bigg/ \sqrt{\sum_{i=1}^{N} D_{A_i}^2 \cdot \sum_{i=1}^{N} D_{B_i}^2} \tag{5}$$

and the other is a measurement we call basic-variation-bias (BVB), which reflects the tempo discrepancy with reference, based on a basic tempo variation.

$$M_{BVB} = \sum_{i=1}^{N}\left|1 - \frac{R_i}{\overline{R}}\right| \quad , \quad R_i = \frac{D_{A_i}}{D_{B_i}} \tag{6}$$

where, $\overline{R}$ refers to the average value of all R_i. The note corresponding relationship in Eq. (5) and Eq. (6) is determined by the alignment result in Sect. 3.1 (see [8] for detail).

3.3 Timbre Brightness

According to previous studies [5], [12], a concentration of power at frequency between 2 kHz to 3 kHz is strongly associated with the perception of timbre brightness, ring and warmth. This could also be explained with the existence of singer's formant, which is described as the prominent spectrum envelope peak near 3 kHz and has been found to greatly affect the sensation of vocal ring [13]. Moreover, some investigations showed that spectral centroid is also a critical feature to the quality of vocal brightness. So in this study, we investigate average energy ratio between 2 kHz to 3 kHz and average spectral centroid as acoustic correspondences of perception "timbre brightness", which are defined as:

$$R_{2k\_3k} = \frac{1}{m}\sum_{i=1}^{m} R_i \quad , \quad R_i = \frac{\int_{2k}^{3k} S_i(f)df}{\int S_i(f)df} \tag{7}$$

and,

$$Cent = \frac{1}{m}\sum_{i=1}^{m} Cent_i \quad , \quad Cent_i = \frac{\int f \cdot S_i(f)df}{\int S_i(f)df} \tag{8}$$

where, $S_i(f)$ refers to the short-time Fourier spectrum of i-th frame at frequency f.

3.4 Vocal Clarity

The perceptual criterion "vocal clarity" is declared to somehow relate to the degree of nonharmonic noise in the vocal tone [5]. So we investigate the average harmonic-to-noise ratio (HNR) measurement as an acoustic correspondence, which can be calculated as:

$$HNR_i = \frac{\sum_j \left(\int_{f_i^j-\sigma}^{f_i^j+\sigma} S_i(f)df \right)}{\int S_i(f)df - \sum_j \left(\int_{f_i^j-\sigma}^{f_i^j+\sigma} S_i(f)df \right)} \tag{9}$$

In addition, we find that the width of harmonic partials has a significant correlation with the feeling of "clarity". So the average narrow-band-to-wide-band energy ratio (NWR) around harmonic partials is also investigated.

$$NWR_i = \frac{\sum_j \left(\int_{f_i^j-\sigma_n}^{f_i^j+\sigma_n} S_i(f)df \right)}{\sum_j \left(\int_{f_i^j-\sigma_w}^{f_i^j+\sigma_w} S_i(f)df \right)} \tag{10}$$

in Eq. (9) and Eq. (10), f_i^j represents the j-th partial frequency of the i-th frame, σ_n and σ_w refer to narrow band width and wide band width respectively. σ in Eq. (9) means the harmonic band width and it is equal to σ_n in our implementation. Finally, $\overline{HNR}$ and $\overline{NWR}$ are used as acoustic correspondences for "vocal clarity".

3.5 Feature Combination

Aiming to find more relevant objective correspondences for subjective perception, a feature combination approach is also investigated in this study. Support Vector Machine (SVM) regression framework is used as a learning tool to mine the feature integration pattern from training data. We use several related acoustic parameters as features and their corresponding perceptual ratings as regression target. Specifically, we combine C_{RA} and C_{DTW} into C_I to relate to intonation accuracy, M_{Cos} and M_{BVB} into C_R to relate to rhythm consistency, R_{2k_3k} and $Cent$ into C_T to relate to timbre brightness, HNR and NWR into C_V to relate to vocal clarity, and finally combine all acoustic features into C_{All} to relate to the overall evaluation.

4. Experiments

4.1 Singing Clips Dataset

The singing samples are recorded from 10 subjects (6 male and 4 female) using normal microphone in 16K-16 bit format. All the subjects have not received any formal vocal training, and their ages are among 20 to 30. The target songs are freely chosen by the singing subjects themselves from a large pop-style song database. Finally, 200 singing clips are used in this study, and most of the clips are about 15 seconds long.

4.2 Evaluation Subjects and Rating Data

Perceptual ratings are obtained from 5 human judges. All the judges have strong music background (two of them are music teachers and the other three are professional musicians). For every singing clip, the judges are asked to assign a rating score from 0 to 10 for every perceptual criterion described in Sect. 2, according to their feelings (0~2 very poor, 2~4 poor, 4~6 so-so, 6~7 well, 7~9 good, 9~10 excellent). The consistency of the judgements was examined and verified by a strong average Spearman rank correlation coefficient of 0.705 (as shown in Table 1, details of the symbols can be found in Sect. 4.3 and Sect. 5).

4.3 Experimental Setting

We investigate the relationship between objective acoustic measurements (including combined features) and perceptual ratings, with Spearman rank correlation coefficient (SRCC) [14], which is defined as:

$$\rho = 1 - \frac{6}{N^3 - N}\sum_{i=1}^{N}(a_i - b_i)^2 \tag{11}$$

where N is the number of components and a_i, b_i are the rank value of the ith component of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Here we take the average score of all judges as the integrated rating data, and the SRCC are measured between acoustic cues and the integrated ratings actually. And all the SVM training and

Table 1 The consistency of the perceptual rating data.

ρ	IA	RC	TB	VC	OP	Mean
Judge 1	0.737	0.650	0.609	0.632	0.755	0.676
Judge 2	0.807	0.694	0.667	0.696	0.792	0.731
Judge 3	0.751	0.655	0.647	0.687	0.715	0.691
Judge 4	0.789	0.697	0.719	0.746	0.768	0.743
Judge 5	0.681	0.619	0.636	0.727	0.738	0.680
Mean	0.753	0.663	0.656	0.698	0.754	**0.705**

Table 2 Spearman rank correlation coefficients (ρ) between perceptual ratings and acoustic measurements.

ρ	C_{RA}	C_{DTW}	C_I	M_{Cos}	M_{BVB}	C_R	R_{2k_3k}	$Cent$	C_T	HNR	NWR	C_V	C_{All}
IA	0.654	0.506	**0.738**	–	–	–	–	–	–	–	–	–	–
RC	–	–	–	0.558	0.359	**0.645**	–	–	–	–	–	–	–
TB	–	–	–	–	–	–	0.260	0.392	**0.682**	–	–	–	–
VC	–	–	–	–	–	–	–	–	–	0.517	0.652	**0.650**	–
OP	0.571	0.422	0.671	0.443	0.332	0.529	0.207	0.319	0.619	0.378	0.486	0.488	**0.758**

testing for the feature combination procedure are based on a 4-fold cross validation criterion.

Moreover, the reliability of the proposed singing evaluation system, which is able to map objective parameters to performance ratings, is examined with two criteria. One is SRCC between mapped performance scores (actually the same as the combined features) and the integrated perceptual ratings. The other is the absolute error criterion which measures the absolute difference between them. Finally the reliability of the objective system is compared to that of human judges.

5. Results and Discussions

Table 2 shows the SRCC between perceptual ratings and acoustic measurements, including combined features. “IA” refers to intonation accuracy, “RC” stands for rhythm consistency, “TB” for timbre brightness, “VC” for vocal clarity and “OP” means overall performance. As can be seen, though correlation between perceptual ratings and some raw acoustic features is not high enough (e.g. M_{BVB} with “RC”, R_{2k_3k} with “TB”), strong correlation can be found between perceptual ratings and the combined features (average ρ of the combined features is 0.694). This demonstrates the critical effect of the feature combination approach. We also think the strong correlation between subjective evaluation and the combined features may illustrate that the original acoustic features grasp complementary parts of the target perception. Moreover, we can find that correlation between “OP” and C_I is relative strong. This result accords well with the common sense that intonation accuracy is the most important aspect of singing and the most crucial criterion for singing evaluation.

Table 3 shows the reliability of the proposed objective evaluation approach, compared to the evaluation reliability of human expert judges. Specifically, we examine the average performance of the objective evaluation system (due to cross validation) by comparing the mapping score to the integrated ratings. And the average performance of the five human judges, by comparing their rating scores to the integrated rating scores. In Table 3, “ME/SD” refers to the mean absolute error and the standard deviation. As we see, the proposed system shows a significant correlation with the integrated ratings (average ρ of 0.694), comparable to human judges (average ρ of 0.705). And the results under absolute error criterion also illustrate that our objective evaluation system (average ME/SD of 0.453/0.372) is completely comparable to human judges (average ME/SD of 0.451/0.349). That is to say, automatic objective singing evaluation from acoustic parameters is possible and is as reliable as expert judges.

Table 3 The reliability test results of the proposed system.

Perceptual Criterion	Objective System		Human Judge	
	SRCC	ME/SD	SRCC	ME/SD
IA	0.738	0.561/0.440	0.753	0.574/0.400
RC	0.645	0.413/0.320	0.663	0.365/0.303
TB	**0.682**	0.393/0.304	0.656	0.445/0.370
VC	0.650	0.438/0.354	0.698	0.424/0.323
OP	**0.758**	0.459/0.440	0.754	0.447/0.351
Mean	0.694	0.453/0.372	0.705	0.451/0.349

6. Conclusion

This letter investigates the relationship between objective acoustic measurements and critical perceptual ratings of singing performance evaluation for untrained singers. In addition to investigations on simple acoustic parameters, we developed a feature combination approach to integrate several relevant features, aiming to find more relevant objective correspondences. Based on these studies, an automatic objective singing performance evaluation approach is proposed and experiment results show that it is in strong correlation with subjective ratings and is as reliable as human judges.

References

[1] J. Wapnick and E. Ekholm, “Expert consensus in solo voice performance evaluation,” J. Voice, vol.11, no.4, pp.429–436, 1997.

[2] J. Oates, B. Bain, P. Davis, J. Chapman, and D. Kenny, “Development of an auditory-perceptual rating instrument for the operatic singing voice,” J. Voice, vol.20, no.1, pp.71–81, 2006.

[3] C. Cao, M. Li, J. Liu, and Y. Yan, “A study on singing performance evaluation criteria for untrained singers,” 9th International Conference on Signal Processing (ICSP08), 2008.

[4] W. Bartholomew, “A physical definition of good voice-quality in the male voice,” J. Acoust. Soc. Am., vol.6, p.25, 1934.

[5] E. Ekholm, G. Papagiannis, and F. Chagnon, “Relating objective measurements to expert evaluation of voice quality in western classical singing: Critical perceptual parameters,” J. Voice, vol.12, no.2, pp.182–196, 1998.

[6] J. Sundberg, “The KTH synthesis of singing,” Advances in Cognitive Psychology, vol.2, no.2-3, pp.131–143, 2006.

[7] T. Nakano, M. Goto, and Y. Hiraga, “An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features,” Ninth International Conference on Spoken Language Processing, 2006.

[8] X. Wu, M. Li, J. Liu, J. Yang, and Y. Yan, “A top-down approach to melody match in pitch contour for query by humming,” Fifth International Symposium on Chinese Spoken Language Processing, 2006.

[9] J. Jang, C. Hsu, and H. Lee, "Continuous HMM and its enhancement for singing/humming query retrieval," Proc. 6th International Conference on Music Information Retrieval (ISMIR 2005), pp.546–551, 2005.

[10] C. Cao, M. Li, J. Liu, and Y. Yan, "Singing melody extraction in polyphonic music by harmonic tracking," Proc. 8th International Conference on Music Information Retrieval (ISMIR 2007), pp.373–374, 2007.

[11] M. Li and Y. Yan, "A humming based approach for music retrieval," J. Voice, vol.21, pp.433–437, 2005.

[12] G. Welch, D. Howard, E. Himonides, and J. Brereton, "Real-time feedback in the singing studio: An innovatory action-research project using new voice technology," Music Education Research, vol.7, no.2, pp.225–249, 2005.

[13] J. Sundberg, The Science of Singing Voice, Northern Illinois University Press, 1987.

[14] M. Kendall, Rank Correlation Methods, Oxford University Press, New York, 1990.