The Evaluation of Singing Voice Accuracy: A Comparison Between Subjective and Objective Methods

*Pauline Larrouy-Maestri, †Yohana Lévêque, ‡Daniele Schön, †Antoine Giovanni, and *Dominique Morsomme,
*Liège, Belgium, †‡Marseille, France

Summary: Objective. Vocal accuracy of a sung performance can be evaluated by two methods: acoustic analyses and subjective judgments. Acoustic analyses have been presented as a more reliable solution but both methods are still used for the evaluation of singing voice accuracy. This article presents a first time direct comparison of these methods.
Methods. One hundred sixty-six untrained singers were asked to sing the popular song "Happy Birthday." These recordings constituted the database analyzed. Acoustic analyses were performed to quantify the pitch interval deviation, number of contour errors, and number of tonality modulations for each recording. Additionally, 18 experts in singing voice or music rated the global pitch accuracy of these performances.
Results. A high correlation occurred between acoustic measurements and subjective rating. The total model of acoustic analyses explained 81% of the variance of the judges' scores. Their rating was influenced by both tonality modulations and pitch interval deviation.
Conclusions. This study highlights the congruence between objective and subjective measurements of vocal accuracy within this first time comparison. Our results confirm the relevance of the pitch interval deviation criterion in vocal accuracy assessment. Furthermore, the number of tonality modulations is also a salient criterion in perceptive rating and should be taken into account in studies using acoustic analyses.
Key Words: Vocal accuracy–Singing–Music experts–Untrained singers–Acoustic analysis–Perceptive rating.

INTRODUCTION

The intonation of a sung melody, in terms of pitch accuracy, is an important factor to determine the singing talent.[1] In experimental psychology, the accuracy of vocal performances has often been evaluated by music experts.[2–5] Vocal accuracy can also be objectively quantified by measuring the fundamental frequency (f_0) variations along the performance. Since the *SINGAD* (SINGing Assessment and Development) system of Howard and Welch,[6] these acoustic methods have been developed[7] and presented as a more reliable solution to evaluate vocal accuracy, as they avoid the natural limits of a subjective judgment such as the imprecision of the exact deviation from the model pitch or the categorization of the pitch information with respect to the closest musical value.[8] Acoustic analysis consists in segmenting the auditory signal and extracting the f_0 of sung vowels. Indeed, vowels carry the maximum of voicing and stable pitch information[9] and mark the onsets of musical tones.[10] In pitch-matching tasks, vocal accuracy is directly represented by the difference between the produced pitch and the model.[5,11–15] In melodic contexts,[16–20] measures are rather based on the relative pitch differences linked to the succession of intervals and so avoid the effect of a change in key at the beginning of a tune, which would lead to errors in the rest of the melody.[4]

Though acoustic analyses have the advantage to provide an objective and reproducible measurement of vocal accuracy,[17,18,20,21] they also have some limits. First, objective measurements require choosing a set of features or dependent variables that possibly describe vocal accuracy and/or what should be considered as an "error."[8] Second, computer-assisted methods seem limited when it comes to vocal accuracy assessment of a full song.[5] As the interpretation is part of the quality of a musical performance, Wise and Sloboda[5] believe that a rating scale suits the evaluation of a melody better than signal processing methods. Third, objective tools for acoustic analysis may not be adapted to all vocal data types. For example, objective measurements adapted for a song performed by untrained singers do not suit when an operatic technique is used.[22] Finally, despite the increase of automated tools, the acoustic analysis methods are still time-consuming and not easily implementable in contexts such as musical or clinical evaluation.

Although technical advances reversed the trend from perceptive methods to acoustic methods in research on singing,[8] the judges' assessment and the objective measurements have never been clearly compared. Otherwise, the comparison between the self-evaluation and the objective measurements is well documented. Studies about tone deafness and poor-pitch singers showed the difficulty for participants to evaluate accurately their musical abilities.[5,14,23]

In the present study, we compared objective and subjective methods assessing the vocal accuracy of a melody sung by 166 participants. Our aim was to (a) estimate the correlation level between the two methods and (b) find out which acoustic measurements can predict the judges' scores. For (a), we used the following variables: a global rating by experts versus a quantitative measurement of pitch interval deviation, the most

Accepted for publication November 13, 2012.
From the *Logopédie de la Voix, Department of Psychology: Cognition and Behaviour, University of Liège, Liège, Belgium; †Laboratoire Parole et Langage, CNRS and Aix-Marseille University, Marseille, France; and the ‡Institut de Neurosciences Cognitives de la Méditerranée, CNRS and Aix-Marseille University, Marseille, France.
Address correspondence and reprint requests to Pauline Larrouy-Maestri, Logopédie de la voix, Département de Psychologie: cognition et comportement, Université de Liège, B-38, Rue de l'Aunaie 30, 4000 Liège, Belgium. E-mail: pauline.larrouy@ulg.ac.be
Journal of Voice, Vol. ■, No. ■, pp. 1-5
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2012.11.003

common acoustic criterion to assess vocal accuracy.[16–18,20,22] For (b), we examined the predictive power of two additional variables: the number of contour errors and the number of tonality modulations. The importance of the melodic contour in music processing has been continually demonstrated using a wide array of perceptual encoding, similarity, and memory paradigms.[24] The second variable has never been directly measured using acoustic methods although some previous studies showed its perceptive relevance for the vocal accuracy evaluation of children or adult performances.[25,26]

METHODS

Participants

One hundred sixty-six untrained singers (57 men and 109 women) were recruited among the Belgian population. Their age ranged from 14 to 76 years ($M = 29.89$ years, standard deviation [SD] = 14.47). The majority (64%) reported listening to music less than 1 hour a day, only 6% declared going to the concerts more than once a month. None of them were professional musicians, 3% had studied a musical instrument in a conservatory of music and 24% reported a basic musical education in a local music school.

The jury comprised 18 subjects with expertise in singing voice and/or music. Four music students ($M = 20$ years) with an average of 12.5 years of musical training were recruited in conservatories and were all following piano classes. The professional musicians were five instrumentalists ($M = 39.8$ years) with an average of 28.2 years of musical training and five singers ($M = 39.4$ years) with an average of 16.6 years of vocal training. They all followed a classical music education in high institutions and were still performing in public when the study took place. The singing voice experts were four speech therapists specialized in singing voice treatment ($M = 31$ years).

Procedure

Song recording. Participants were asked to produce two vocal glissandi (sliding up to the highest pitch and down to the lowest pitch in a comfortable range). The experimenter illustrated the vocal exercise by singing and imitating manually the movement of the voice. The aim of these glissandi was to warm up the vocal organs,[27] verify the vocal capacity of the subjects, and encourage a lack of inhibition in front of the experimenter and the recording equipment. Then, they performed individually the popular French song "Happy Birthday" a cappella. Participants were instructed to sing "naturally, while imagining a festive and friendly context." No particular starting note was given to let the participant choose his/her comfortable range. The sound recordings were made using a head-worn microphone (Sennheiser HS2, Wedemark, Germany) positioned at a constant distance of 2 cm from the right corner of the mouth and a Marantz Professional Solid State Recorder (PMD67; Marantz, Kanagawa, Japan). All these sung performances are listed in a database, which can be viewed through the following link: http://sldr.org/sldr000774/en.

Self-evaluation. After the sung performance, participants were instructed to rate their proficiency to sing in tune on a nine-point scale with 1 indicating "very inaccurate" and 9 "very accurate."

Judge's rating. Each jury member listened to the 166 recorded songs through headphones and rated them on a nine-point scale with 1 indicating "very inaccurate" and 9 "very accurate." The instruction was to take into account the overall vocal pitch accuracy to make their judgments and no other criteria than pitch. Four practice trials selected from the database were presented to verify the understanding of the instruction and allow the judges to adjust their assessment. They performed the evaluation individually.

Description of the objective method

Acoustic analysis. The popular song "Happy Birthday" comprised four phrases. It has 25 notes (each note corresponding to a syllable) and 21 when one ignores the rhythmically concurrent repeated notes (Figure 1). Data processing was semi-automatically done in two stages on a MacBook Pro (*Mac OS X*, Version 10.6.5; Apple, Cupertino, CA). Analyses were performed through *AudioSculpt 2.9.4v3* and *OpenMusic 6.3* softwares (IRCAM, Paris, France) using a short-time Fourier transform analysis. Markers were manually placed on the spectrogram, where the f_0 and the 10 first partials were clearly visible, to avoid the attacks and the glides between notes. Then, the mean f_0 estimation in the segments was automatically calculated and converted into cents (1/2 tone = 100 cents) for the 21 notes of the tune.

Criteria observed. The measurements are based on the melodic intervals. The calculations are grounded on the equal temperament (ie, constant frequency multiple between the notes of the chromatic scale), which is a compromise tuning scheme used in Western music. As Figure 2 illustrates, for each production, three criteria of vocal accuracy were quantified: pitch interval deviation, number of contour errors, and tonality modulations.

Pitch interval deviation. We calculated the difference in cents between each performed interval and the theoretical one. We considered the absolute value of the differences (to avoid sharp and flat errors canceling each other) and computed the average score across the entire melody. A small deviation reflects a high precision of intervals.

Contour errors. We counted each time that the produced interval direction deviated from the direction of the musical score.

Tonality modulation. We computed the number of modulations, defined as an interval error larger than 100 cents not followed by a corrective interval of at least 100 cents in the reverse

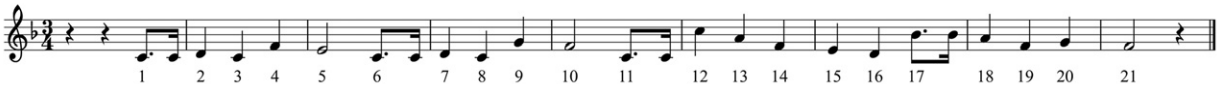

FIGURE 1. Score of the tune "Happy Birthday" with the number of notes used for calculating accuracy.

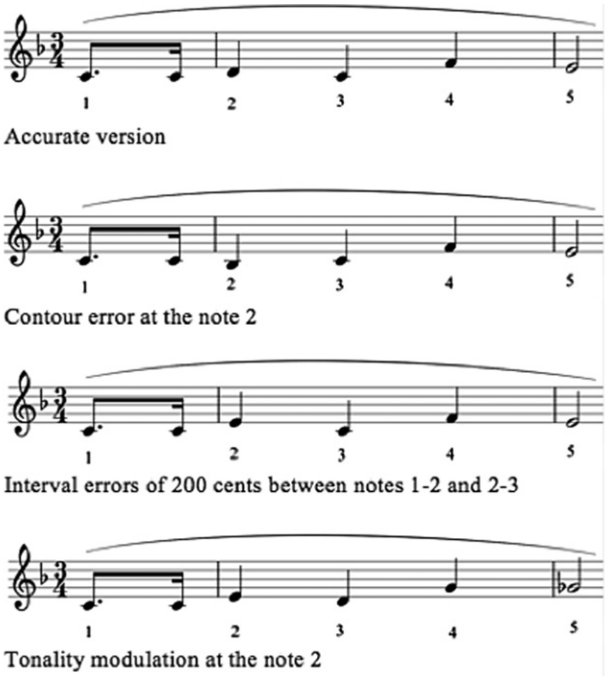
1
2
3
4
5
Accurate version
1
2
3
4
5
Contour error at the note 2
1
2
3
4
5
Interval errors of 200 cents between notes 1-2 and 2-3
1
2
3
4
5
Tonality modulation at the note 2

FIGURE 2. Illustration of the errors observed through the objective method.

direction. Thus, modulations indicated that tonality changed during at least three notes (two intervals).

RESULTS

We computed a correlation matrix using Spearman coefficient to estimate the pairwise correlations between the 18 judges. We found a median correlation coefficient of $r = 0.77$ (SD = 0.08, $P < 0.01$). As the 18 judges provided strongly and significantly correlated ratings, the mean rating from the whole group has been used in the following analyses.

Interval deviation and subjective judgments of vocal accuracy

Deviation from target intervals is the most common criterion used in acoustic analyses to estimate the vocal accuracy. To know to what extent this criterion is correlated with subjective judgments, we calculated correlations between interval deviation, scores of vocal accuracy given by the judges, and self-evaluation of vocal accuracy using the Spearman coefficient for the 166 subjects.

We found a high and significant correlation between the interval deviation criterion and the average score given by the judges ($r = -0.87$; $P < 0.01$), as illustrated in Figure 3 (top).

We then computed this correlation coefficient with subgroups of judges of different size (N = 1–16). For this purpose, we took random subgroups among the original 18 judges (100 random subgroups of each size, from 1 to 16 judges) and computed the correlation between the mean rating of each subgroup and the semi-automatic interval deviation measure. We found a slight decrease of r for smaller groups of judges. Figure 4 shows the relationship between the correlation coefficient r and the number of judges. Data suggest that three judges are sufficient to get an average correlation coefficient superior to 0.85.

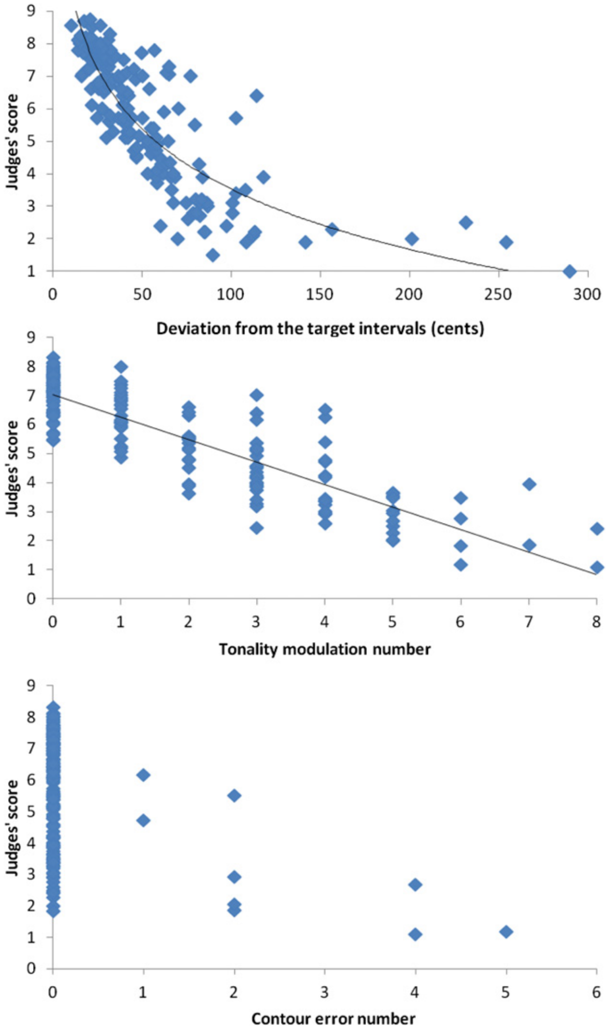

FIGURE 3. Relationship between judges' scores and interval deviation (top), judges' scores and tonality modulation number (middle), and judges' scores and contour error number (bottom).

Self-evaluation by the subjects was moderately but significantly correlated to the interval deviation criterion, $r = -0.35$; $P < 0.01$ and in the same extent to the judges' scores, $r = -0.35$; $P < 0.01$. Self-evaluation appeared affected by some overrating. Notably, 41% of the inaccurate singers (judges' score<4, N = 39) overestimated their vocal skill (self-evaluation>4). Conversely, a larger number of participants undervalued their vocal skill: 63% of the very accurate singers (judges' score>6, N = 84) undervalued themselves (self-evaluation<6).

Acoustic parameters predicting the judge rating

To determine whether acoustic variables other than the interval deviation criterion might contribute to predict the score of vocal accuracy given by the judges, we performed a multiple linear regression analysis entering interval deviation, number of tonality modulations, and number of melodic contour errors. As we observed a logarithmic relationship between judges' scores and

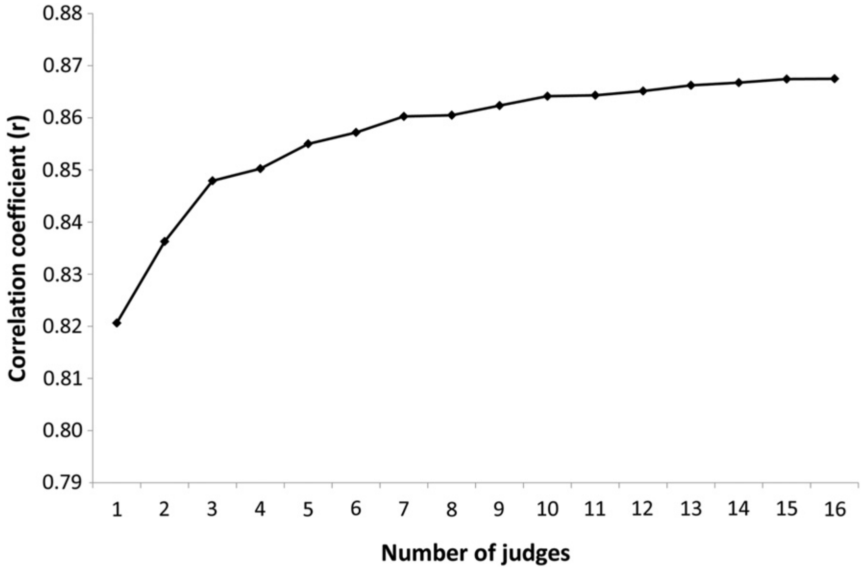

FIGURE 4. Evolution of the correlation interval deviation/average judge score as a function of the number of judges (n) in the panel. We randomly selected 100 subgroups of n judges among our 18 experts and iterated a Spearman correlation test between each subgroup of judges' mean rating and the interval deviation of singers. N varies between 16 and 1.

interval deviation criterion (Figure 3, top), the logarithm of the interval deviation was used in the multiple regression.

The results of the analysis are shown in Table, including beta weights and significance tests for each variable. In addition, The Table displays the results of the significance test for the multiple R^2 for the set of three predictor variables.

As shown in the Table, the result of the regression analysis indicated that two variables predicted the score of vocal accuracy given by the judges: the pitch interval deviation ($\beta = .51$; $t = -6.61$; $P < 0.001$) and the number of tonality modulations ($\beta = -.45$; $t = -6.33$; $P < 0.001$). The number of contour errors did not contribute to explain the judges' scores ($\beta = .08$; $t = 1.89$; $P = 0.06$), probably because this type of error was very rare. The correlation between these two former criteria and the judges' score is illustrated in Figure 3 (middle and bottom).

The total model explained 81% of the variance of judges' scores ($F = 232.17$; $P < 0.001$).

TABLE.
Summary of Multiple Regression Analysis on Judges' Scores With the Three Acoustic Variables Used as Predictors

R^2		0.81	
Adjusted R^2		0.81	P
F		232.17	2.00E-58
	β	t Value	Pr (>|t|)
(Intercept)		16.30	5.68E-36
Intervals	−.51	−6.61	5.43E-10
Tonality	−.45	−6.32	2.39E-09
Contours	.08	1.89	0.06

Intervals, mean size of interval deviation; tonality, number of tonality modulations; contours, number of contour errors. For each variable, the beta weights and significance tests are represented.

DISCUSSION

For the first time, this study directly compared two different methods to assess the vocal accuracy over the same material. We observed that the 18 judges provided significantly correlated ratings in the context of a popular song assessment. This result shows that the voice and/or music experts used similar subjective criteria despite their different backgrounds.

For the 166 sung performances, we found a high and significant correlation between the interval deviation and the scores given by the judges ($r = 0.87$, $P < 0.01$). Previous researches used objective or subjective methods, without comparing them. In this study, both objective and subjective approaches were used and compared, showing similar results in the assessment of vocal accuracy in a melodic context. Interestingly, this correlation between objective and subjective methods was robust enough to survive downsampling to a number of three raters from the judge group. We observed a logarithmic relationship between the pitch interval deviation criterion and the subjective measurement, which means that the perceptual rating was more discriminatory among the accurate singers than for the singers with a low score according to the pitch interval deviation criterion. Judges seemed more critical when an accurate singer made a false note but did not make a difference between slightly inaccurate to very inaccurate sung performances.

When the perceptive assessment was done by the participants themselves, the correlation between interval deviation and subjective measurement was moderate ($r = -0.35$; $P < 0.01$). The present study confirms the difficulties in self-evaluation pointed out in previous studies.[14,23] These results could be attributed to a lack of musical expertise of the participants (eg, a wrong representation of the target melody) or to an intrinsic difficulty of the self-evaluation procedure (eg, the social nature of self-assessments in general or specific auditory feedback integration). Comparing the subjective assessment of the present study with the rating of naive judges, without any music or vocal expertise, would allow a better understanding of the causes of such difficulties. Indeed, a similar pattern of results would indicate that difficulties in self-evaluation would not be due to a lack of musical expertise.

This study shows the correlation between objective and subjective measurements but also allows us to examine the predicting weight of the three criteria used in objective measurements. The regression analysis indicated that the pitch interval deviation criterion was related to the score of vocal accuracy given by the judges ($\beta = .51$; $t = -6.61$; $P < 0.001$). This study thus supports the use of this criterion for the assessment of vocal accuracy in a melodic context.[16–18,20,22] In our study, two other criteria were observed: the number of contour errors and the number of tonality modulations. The regression analysis indicated that the number of contour errors did not explain the judges' scores ($\beta = .08$; $t = 1.89$; $P = 0.06$). This pattern can be discussed regarding the choice of the melody, which leads to few contour errors. This tune has been chosen for its strong tonal center and the diversity of intervals. Also, the French version of the popular song "Happy Birthday" is very common and learned early in the childhood. Years of familiarity (ear and vocal) and the simplicity of the material could explain that untrained singers sang the good contour of the melody. It would be interesting to propose a more complex tune, which would induce more contour errors to be analyzed. Finally, this study highlights that the judges' score is partly explained by the number of tonality modulations ($\beta = -.45$; $t = -6.33$; $P < 0.001$). This variable has never been included in an acoustic analysis so far but this result supports the importance of the tonality component in vocal accuracy assessment.[25,26]

Finally, even if the judges were asked to focus on the pitch of the melody, the unexplained 19% of variance in the regression model could be linked to other criteria such as rhythm accuracy or vocal quality, criteria that could unintentionally intervene in the accuracy judgment. Future research would need to investigate these criteria to better understand the perceptual judgment process.

CONCLUSION

In this study, we compared an objective and a subjective method to assess the vocal accuracy of a popular song performed by 166 untrained singers. Our results highlight the congruence between objective and subjective measures, whereas the subjective ratings are performed by expert judges. In analytical computer-assisted methods, the measured variables can be controlled but their use must be preceded by theoretical choices. Our results clearly confirm the weight of the pitch interval deviation criterion in the vocal accuracy assessment. Furthermore, this study also underlines that the number of tonality modulations is a relevant criterion in perceptive rating and should be taken into account for the objective vocal accuracy assessment.

Acknowledgments

The authors thank the "Centre Henri Pousseur" of Liège and Guillaume Videlier for the technical support.

REFERENCES

1. Watts C, Barnes-Burroughs K, Andrianopoulos M, Carr M. Potential factors related to untrained singing talent: a survey of singing pedagogues. *J Voice*. 2003;17:298–307.
2. Hébert S, Racette A, Gagnon L, Peretz I. Revisiting the dissociation between singing and speaking in expressive aphasia. *Brain*. 2003;126: 1838–1850.
3. Racette A, Bard C, Peretz I. Making non-fluent aphasics speak: sing along! *Brain*. 2006;129:2571–2584.
4. Schön D, Lorber B, Spacal M, Semenza C. A selective deficit in the production of exact musical intervals following right-hemisphere damage. *Cogn Neuropsychol*. 2004;21:773–784.
5. Wise KJ, Sloboda JA. Establishing an empirical profile of self-defined "tone deafness": perception, singing performance and self-assessment. *Music Sci*. 2008;12:3–26.
6. Howard DM, Welch GF. Microcomputer-based singing ability assessment and development. *Appl Acoust*. 1989;27:89–102.
7. Elmer SS, Elmer FJ. A new method for analysing and representing singing. *Psychol Music*. 2000;28:23–42.
8. Dalla Bella S, Berkowska M, Sowinski J. Disorders of pitch production in tone deafness. *Front Psychol*. 2011;2:164.
9. Murayama J, Kashiwagi T, Kashiwagi A, Mimura M. Impaired pitch production and preserved rhythm production in a right brain-damaged patient with amusia. *Brain Cogn*. 2004;56:36–42.
10. Sundberg J, Bauerhuppmann J. When does a sung tone start? *J Voice*. 2007; 21:285–293.
11. Alcock KJ, Passingham RE, Watkins K, Vargha-Khadem F. Pitch and timing abilities in inherited speech and language impairment. *Brain Lang*. 2000;75:34–46.
12. Alcock KJ, Wade D, Anslow P, Passingham RE. Pitch and timing abilities in adult left-hemisphere-dysphasic and right-hemisphere-damaged subjects. *Brain Lang*. 2000;75:47–65.
13. Amir O, Amir N, Kishon-Rabin L. The effect of superior auditory skills on vocal accuracy. *J Acoust Soc Am*. 2003;113:1102–1108.
14. Pfordresher PQ, Brown S. Poor-pitch singing in the absence of "tone deafness". *Music Percept*. 2007;25:95–115.
15. Watts C, Moore R, McCaghren K. The relationship between vocal pitch-matching skills and pitch discrimination skills in untrained accurate and inaccurate singers. *J Voice*. 2005;19:534–543.
16. Berkowska M, Dalla Bella S. Reducing linguistic information enhances singing proficiency in untrained singers. *Ann N Y Acad Sci*. 2009;1169:108–111.
17. Dalla Bella S, Berkowska M. Singing proficiency in the majority. *Ann N Y Acad Sci*. 2009;1169:99–107.
18. Dalla Bella S, Miguère JF, Peretz I. Singing proficiency in the general population. *J Acoust Soc Am*. 2007;121:1182–1189.
19. Pfordresher PQ, Brown S. Enhanced production and perception of musical pitch in tone language speakers. *Atten Percept Psychophys*. 2009;71: 1385–1398.
20. Pfordresher PQ, Brown S, Meier KM, Belyk M, Liotti M. Imprecise singing is widespread. *J Acoust Soc Am*. 2010;128:2182–2190.
21. Terao Y, Mizuno T, Shindoh M, et al. Vocal amusia in a professional tango singer due to a right superior temporal cortex infarction. *Neuropsychologia*. 2006;44:479–488.
22. Larrouy-Maestri P, Morsomme D. Criteria and tools for objectively analysing the vocal accuracy of a popular song. *Logoped Phoniatr Vocol*. 2012; [Epub ahead of print].
23. Cuddy LL, Balkwill LL, Peretz I, Holden RR. Musical difficulties are rare: a study of "tone deafness" among university students. *Ann N Y Acad Sci*. 2005;1060:311–324.
24. Schmuckler MA. Components of melodic processing. In: Hallam, ed. *The Oxford Handbook of Music Psychology*. Oxford, United Kingdom: Oxford University Press; 2009:93–106.
25. Flowers PJ, Dunne-Sousa D. Pitch-pattern accuracy, tonality, and vocal range in preschool children's singing. *J Res Music Educ*. 1990;38:102–114.
26. Price HE. Interval matching by undergraduate nonmusic majors. *J Res Music Educ*. 2000;48:360–372.
27. Amir O, Amir N, Michaeli O. Evaluating the influence of warmup on singing voice quality using acoustic measures. *J Voice*. 2005;19:252–260.