The Evaluation of Vocal Pitch Accuracy:

The Case of Operatic Singing Voices

Pauline Larrouy-Maestri[1], David Magis[2], & Dominique Morsomme[1]

1. Department of Psychology: Cognition and Behavior, University of Liège, Belgium

2. Department of Education, University of Liège, Belgium.

Abstract

The objective analysis of Western operatic singing voices indicates that professional singers can be particularly "out of tune". This study aims to better understand the evaluation of operatic voices, which have particularly complex acoustical signals. Twenty-two music experts were asked to evaluate the vocal pitch accuracy of 14 sung performances with a pairwise comparison paradigm, in a test and a retest. In addition to the objective measurement of pitch accuracy (pitch interval deviation), several performance parameters (average tempo, fundamental frequency of the starting note) and quality parameters (energy distribution, vibrato rate and extent) were observed and compared to the judges' perceptual rating. The results show high intra- and inter-judge reliability when rating the pitch accuracy of operatic singing voices. Surprisingly, all the parameters were significantly related to the ratings and explain 78.8% of the variability of the judges' rating. The pitch accuracy evaluation of operatic voices is thus not based exclusively on the precision of performed music intervals but on a complex combination of performance and quality parameters.

Keywords: acoustical analyses, music experts, operatic technique, singing, intonation

Introduction

In classical music, singing in tune is critical, as it enables singers to create expressive melodic lines and harmonize with other performers. In addition, precision of intonation has been shown to be an important factor associated with the perception of singing talent (Watts, Barnes-Burroughs, Andrianopoulos, & Carr, 2003). Accuracy of vocal pitch, in terms of the precision of performed music intervals, can be objectively measured by observing fundamental frequency variations over the course of the tune (Berkowska & Dalla Bella, 2009; Dalla Bella & Berkowska, 2009; Dalla Bella, Giguère, & Peretz, 2007; Larrouy-Maestri & Morsomme, in press; Pfordresher & Brown, 2009; Pfordresher, Brown, Meier, Belyk, & Liotti, 2010). A previous study found that objective and subjective measurements of the vocal pitch accuracy of untrained singers are highly correlated (Larrouy-Maestri, Lévêque, Schön, Giovanni, & Morsomme, 2013). That study also reported that the objective measurements of pitch accuracy explained 81% of the variance in the judges' ratings.

Interestingly, renowned professional singers are not always found to sing in tune when their performances are objectively measured (Sundberg, Prame, & Iwarsson, 1996; Vurma & Ross, 2006). Recently, Larrouy-Maestri and Morsomme (in press) compared 63 untrained and 14 trained singers' performance of the popular song "Happy Birthday" with a computer-assisted method. The trained singers had to perform the melody twice, with and without a Western operatic singing technique. As expected, the professional singers were more accurate than the untrained ones. Intriguingly, an effect of vocal technique appeared: music intervals were less precise when a Western operatic singing technique was employed than when it was not. This pattern was confirmed with a larger group of professional

singers across two different melodies (Larrouy-Maestri, Magis, & Morsomme, in press). Indeed, the pitch interval deviation was greater with the operatic singing technique than without it, no matter which melody was performed (a popular song or a melody from the romantic period). Given this negative impact of operatic technique on performed pitch accuracy, and that professional musicians are expected to sing in tune (and thus to respect the size of the intervals), this study aims to examine the perceptual judgment of the pitch accuracy of operatic singing voices.

The operatic singing technique influences several acoustical components such as vibrato and reinforcement of the energy band containing the singer's formant (Sundberg, 2013). Vibrato has been extensively investigated since Seashore (1938) and has often been observed among classical singers (e.g., Bretos & Sundberg, 2003; De Almeidera, Cukier-Blaj, Duprat, Camargo, & Granato, 2009; Ekholm, Papagiannis, & Chagnon, 1998; Hirano, Hibi, & Hagino, 1995; Howes, Callaghan, Davis, Kenny, & Thorpe, 2004; Prame, 1994, 1997; Stone, Cleveland, Sundberg, & Prokop, 2002; Sundberg, 1994). It corresponds to a quasi-periodic modulation of a tone's frequency that can be further characterized by its rate (number of fluctuations per second) and extent (amplitude of fluctuations around the pitch of a tone). Regarding the spectral distribution, the operatic singing technique is associated with an increase in energy between 2 and 4 kHz (Barnes, Davis, Oates, & Chapman, 2004; Omori, Kacker, Carroll, Riley, & Blaugrund, 1996; Sundberg, 1995, 2001, 2013; Thorpe, Cala, Chapman, & Davis, 2001), which allows the voice to be heard above the sound of the orchestra. These two components (i.e., vibrato and reinforcement of the energy band containing the singer's formant) seem to be salient characteristics listeners are highly sensitive to (e.g. Ekholm et al., 1998; Garnier, Henrich, Castellengo, Sotiropoulos, &

Dubois, 2007). For instance, Garnier et al. (2007) proposed a psycholinguistic study of the singing teachers' verbal descriptions of voice quality. From a precise linguistic analysis, the authors explored the verbal descriptors provided by singing teachers to describe vocal performance in Western operatic singing. They pointed out that vibrato and timbre are relevant and consistent descriptors of a "good" vocal performance.

However, the acoustical features developed in operatic singing techniques can affect how the pitch height of the voice is perceived. Indeed, pitch perception of tones differs depending on spectral composition (Hutchins, Roquet, & Peretz, 2012; Russo & Thompson, 2005; Vurma, Raju, & Kuuda, 2010; Warrier & Zatorre, 2002) and vibrato (van Besouw, Brereton, & Howard, 2008). As the evaluation of performed pitch intervals is based on the pitch level of individual notes and the perception of tones is altered by the acoustical components involved (and appreciated) in operatic singing voices, it may be particularly complex to evaluate such voices. The consequence of this complexity has been observed in expert listeners' unexpected tolerance of both "in tune" and "out of tune" intervals (Vurma & Ross, 2006). In Vurma and Ross's study, 17 expert listeners were asked to rate isolated intervals performed by trained singers. The authors remarked that the judges sometimes considered intervals as "in tune" even when they were off by 20 to 25 cents. Sundberg et al. (1996) were interested in judgment of pitch accuracy within a melodic context. Seven listeners had to circle each tone they perceived to be "out of tune" in a set of 10 commercial recordings (Schubert, *Ave Maria*). When analyzing the mean F_0 of both circled and noncircled tones, the authors found considerable variability among the judges and an overall tolerance zone of about 10 cents (wider for some tones). Although these studies highlighted the tolerance of music experts regarding professional singers' pitch

accuracy, the experimental designs were not sensitive to the evaluation process used by the music experts. In addition, the cause of this tolerance could be observed by analyzing and comparing the acoustical parameters involved in the operatic singing technique.

The present study aims to (a) assess the ability of expert listeners to evaluate the vocal pitch accuracy of operatic singing voices, and (b) find out which acoustical measurements can predict the evaluation of pitch accuracy in operatic singing voices. In pursuit of goal (a), we observed intra- and inter-judge reliability in assessing the overall pitch accuracy of operatic singing voices and hypothesized that high intra-judge reliability would support the ability of expert listeners to rate the overall pitch accuracy of complex voices. In addition, the observation of inter-judge reliability would allow us to investigate consistency between the judges. For goal (b), we objectively measured several acoustical features and incorporated them in a predictive model of vocal pitch accuracy evaluation. Doing so would hopefully illuminate upon the process of subjective evaluation of pitch accuracy in operatic singing voices.

Method

Participants

The judges were 22 professional musicians (8 women, 14 men) aged 26 to 73 years old (M = 45.68 years, SD = 11.16). They had between 15 and 55 years of music experience (M = 35.77 years, SD = 10.74) and all received a classical music education (all instrument families were represented) in higher institutions such as music conservatories. When the study took place, all participants were still performing in public and reported practicing their instrument(s) 18.68 hours/week on average.

Material

All the sung performances used in the perceptual task were selected from the database: http://sldr.org/sldr000792/en. Professional or semi-professional singers were asked to produce two vocal glissandi. Then they individually performed, a cappella, two different melodies (the French version of "Happy Birthday" and a romantic melody of their choice) with two different techniques (singing without any particular technique and with a Western operatic singing technique). For the popular song, no particular starting note was given so the participants could perform in their comfortable range. In this study, we examined the last sung performance recorded, namely the popular song "Happy Birthday" performed with an operatic singing technique. We therefore presumed that the singers were warmed up and were comfortable enough to perform a characteristic sample of operatic singing technique.

The sound recordings were made using a head-worn microphone (Sennheiser HS2, Wedemark, Germany) positioned at a constant distance of 2 cm from the left corner of the mouth as well as a Marantz Professional Solid State Recorder PMD67 (Kanagawa, Japan). The performances were recorded with a sampling frequency of 44.1 kHz and a resolution of 16 bits.

In order to apply a pairwise comparison paradigm, we needed relatively short stimuli and thus extracted the last sentence of "Happy Birthday" (Figure 1). From the initial corpus, we selected performances by females (from 245.42 Hz to 449.26 Hz, M = 352.55 Hz, SD = 21.13) in order to limit the recognition of the music excerpts by the judges and to avoid an eventual bias due to the gender of the singers. In addition, we have chosen the performances in which the last tone of the tune was long enough to examine its vocal

quality (from 1.13 s to 1.98 s, *M* = 1.45 s, *SD* = 0.09). Based on these two conditions, we selected 14 excerpts sung by 14 different professional or semi-professional singers, aged 21 to 66 years old (*M* = 34.86; *SD* = 12.72). They all reported having received an education in classical music as well as delivering regular solo vocal performances in a classical style. They had begun singing lessons between 10 and 49 years old (*M* = 20.79; *SD* = 9.23) and reported between 6 and 17 years of singing lessons (*M* = 10.57; *SD* = 3.58). At the time of the experiment, they reported singing 14.07 hours/week on average.

Acoustical Analysis

For the 14 sung performances, six variables were observed, grouped into performance parameters (pitch interval deviation, average tempo and fundamental frequency of the starting note) and quality parameters (energy distribution, vibrato rate and vibrato extent). All the acoustical analyses were run on a Macintosh (Mac OS X, Version 10.6.8).

Performance parameters. The vocal pitch accuracy measurement was based on the relative pitch differences (F_0 variations) between the 5 tones analyzed in the music excerpt (in cents). Note that this measurement was computed on the basis of an equal temperament, which is a compromise tuning scheme developed for keyboard instruments. Although the equal tempered system may not seem to be appropriate for a cappella choirs (Howard, 2007a, 2007b), it remains a referent in Western tonal music and is commonly used in researches on accuracy of vocal pitch (Berkowska & Dalla Bella, 2009; Dalla Bella & Berkowska, 2009; Dalla Bella et al., 2007; Larrouy-Maestri & Morsomme, in press; Pfordresher & Brown, 2009; Pfordresher et al., 2010). Using AudioSculpt 2.9.4v3 software (IRCAM, Paris, France), a Short-Time Fourier Transform (STFT) analysis was performed

and markers were manually placed on the spectrogram in order to select the stable part of each tone (i.e., avoiding the attacks and the glides between tones). The mean F_0 of each tone was then extracted with OpenMusic 6.3 software (IRCAM, Paris, France). Note that when a vibrato occurred (mainly for the last tone), we selected complete cycles for the mean F_0 measurement. We then calculated the size in cents of each interval between tones 1 and 5 (Figure 1). Each value (i.e. intervals 1-2, 2-3, 3-4 and 4-5) was compared to the standard given by the music score. For instance, the last interval (between tones 4 and 5) should be two semitones, corresponding to 200 cents. If the singer produced an interval of 150 or 250 cents, that would mean a deviation of 50 cents. We then averaged each deviation to generate the pitch interval deviation criterion. A small deviation indicated that the music intervals were very precise.

Insert Figure 1 about here

As performance parameters, we also observed the *fundamental frequency (F_0)* of the starting tone (in Hertz) and the average *tempo*, computed on the basis of the excerpt's length, by dividing the number of beats (i.e. quarter note) contained in the excerpt by the duration (in beats per minute).

Quality parameters. For these acoustical analyses, we extracted the last tone in each melody, which is also the longest tone of the sung excerpt (tone 5 in Figure 1), sung on the vowel */ɛ/*. Note that the acoustic signal was manually segmented with respect to vibrato cycles.

Energy distribution allows one to observe the spectral balance and thus the energy in the band containing the singer's formant. This measurement was computed in two steps, with AudioSculpt and OpenMusic software (IRCAM, Paris, France). First, we selected three

bands from the total frequency range, 0–2.4 kHz, 2.4–5.4 kHz and 5.4–10 kHz (bands commonly selected for analyzing performances by females). Then, the energy of the second band was divided by the energy across all three bands. Note that a high score for the energy distribution variable shows a strong reinforcement of the band containing the singer's formant.

The *vibrato rate* (VR) and *vibrato extent* (VE) were also measured with AudioSculpt and OpenMusic software (IRCAM, Paris, France). These measurements were supported by the fact that vibrato is defined as a quasi-periodic modulation of fundamental frequency, which is generally found to be closely sinusoidal (e.g., Prame, 1994, 1997). For VR, the number of quasi-periodic modulations of the F_0 per second was reported in Hertz. For VE (i.e. amplitude of the F_0 variations within the same tone), the extreme F0 values were measured from the F_0 curve. The difference between the minimum and the maximum was reported in cents.

Perceptual Task

The 14 sung performances were presented with a pairwise comparison paradigm. Each one was compared to the other ones in a random order. Thus, the total number of pairs was N(N–1) / 2, or 91 pairs to compare in the present study.

The participants pushed buttons on a graphical user interface to listen to the two sung performances as often as they wished. For each pair, they were asked to indicate which excerpt was the most "in tune," considering only pitch accuracy as a parameter (and not the quality of the voice). Participants also had the possibility of labeling both excerpts in a pair as "equal" regarding the pitch accuracy criterion. After each pair, they saved their choice and a new trial was presented, that is, another pair of sung performances to compare.

For each judge, the perceptual task was scored in three steps. All sung performances were initialized to a score of zero. For each pair to be compared, the total score of the sung performance considered as the most "in tune" was increased by one point (no points were awarded to the "out of tune" excerpt in the pair). If both sung performances of the pair were judged to be "equal," the total score of both sung performances was increased by 0.5 points. After the presentation of the 91 pairs, the total score of each music excerpt was computed (i.e. accumulation of points over trials). The result was a ranking of the 14 singers for each judge, with higher scores for the sung performances considered to be the most "in tune".

The auditory stimuli were presented through professional headphones (K271 MKII, AKG, Vienna, Austria) at a comfortable loudness level. The entire perception task was approximately 25 minutes long, and participants were given a break halfway through.

The same procedure was administered twice (test-retest) with 8 to 15 days in between, in order to observe the intra-judge reliability.

Statistical Analyses

Intra-judge reliability was evaluated as follows. For each judge, the rankings of the singers were derived from that judge's ratings, separately in the test and the retest. Differences in rankings (test minus retest) were obtained, and the sample variance of the 14 differences in rankings was derived as a measure of intra-judge reliability. Small variances indicated minor changes in rank between the test and the retest, and hence strong intra-judge reliability, while larger variances reflected larger differences in rank and lower intra-judge reliability.

Inter-judge reliability was evaluated in the same way as intra-judge reliability. Each singer was assigned a set of rankings, with each one corresponding to her performance

(compared to the other singers) for one judge. The variance of the rankings was then computed. Small variances indicated small differences in the ranking of singers' performances across judges (and therefore strong inter-judge reliability), while larger variances indicated lower agreement among the judges' rankings.

In practice, issues involving intra- and inter-judge reliability were detected by a similar process. First, the sample distributions of the respective variances were compared to a chi-squared distribution by selecting the degrees of freedom that maximized the p-value of a Kolmogorov-Smirnov test. Then, a threshold was set as the quantile of the chi-squared distribution with a lower tail probability of 0.95. Judges whose variance of differences in rankings (test minus retest) was greater than that threshold were considered to have been affected by an intra-judge reliability problem and were excluded from further analyses.

Finally, the relationships between the judges' ratings and the objective measurements (performance and quality parameters) were investigated through a multiple linear modeling approach. Seven covariates (judge, pitch interval deviation, tempo, F_0, energy, VR and VE) were included as main effects, together with the twenty-one possible pairwise interactions. The model was then simplified to retain the most statistically significant terms while retaining an acceptable R^2 coefficient. The final model was further interpreted in terms of the significant main effects and interactions (through an analysis of variance), as well as the direction and magnitude of meaningful terms.

Results

Intra-Judge Reliability

The distribution of the 22 variances of rank differences is displayed in Figure 2 in a histogram. Because the nature of the variance is an index of intra-judge reliability, an

adjustment of the sample distribution by a chi-squared distribution is expected. The best adjustment is obtained by considering six degrees of freedom, for which the Kolmogorov-Smirnov test returns a p-value of 0.535. The corresponding chi-squared distribution is shown in Figure 2 by a dashed line.

Insert Figure 2 about here

The quantile of this chi-squared distribution with six degrees of freedom and lower tail probability of 0.95 equals 12.592 and serves as a detection threshold for intra-judge reliability issues. Figure 3 presents the 22 variances of rank differences for each judge, represented by their respective numbers. The horizontal dashed line represents the detection threshold.

Insert Figure 3 about here

As Figure 3 shows, judges 2 and 12 exhibited a sample variance larger than the threshold, meaning that their ranking differences were too dispersed with respect to the overall set of variances. In other words, the differences in performance ranks (between the test and the retest) varied more for these two judges compared to the overall variability of these rank differences. Judges 2 and 12 were therefore discarded from further analyses.

Inter-Judge Reliability

Figure 4 is a histogram showing the variances of ranks for each singer's performance. Recall that judges 2 and 12 were removed from this analysis because of their low intra-judge reliability level. The adjustment by a chi-squared distribution is best when seven degrees of freedom are chosen (Kolmogorov-Smirnov $p = 0.924$).

Insert Figure 4 about here

The corresponding quantile with a lower tail probability of 0.95 equals 14.067. As Figure 5 shows, none of the variances of singers' ranks are greater than that threshold, which indicates quite strong inter-judge reliability across all the singers' performances and suggests the use of similar judgment process by all 20 judges.

Insert Figure 5 about here

Predictive Model of Vocal Pitch Accuracy Evaluation

The multiple regression model including all the aforementioned covariates (means of the six objective parameters are presented in Table 1) and main effects has an R^2 coefficient of 82.7%. The final model retained has an R^2 coefficient of 78.8%. This model contains all main effects as well as several pairwise interaction terms. The output of the analysis of variance (ANOVA) is displayed in Table 2.

All covariates are present as main effects (top part of Table 2), meaning that all factors are necessary to explain the judges' ratings of vocal pitch accuracy. However, there is no simple relationship between the judges' ratings and those factors. Indeed, each factor is at least present in one pairwise interaction term, meaning that the effect of this factor on the judges' ratings is not independent from any other factor. This can be seen from the quite long list of significant pairwise interaction effects (bottom of Table 2). Note that some interaction terms were kept in the final model although the Table 2 lists them as nonsignificant interactions. The reason is that, although not statistically significant, their contribution to the overall R^2 coefficient cannot be neglected. This is probably a sign of collinearity between the covariates.

Insert Table 1 about here

Insert table 2 about here

Scores cannot therefore be explained by a single covariate, or by one covariate that does not interact with the others. For instance, the pitch interval deviation parameter does not explain the judges' rating as a main effect but in association with, on the one hand, F_0 of the starting tone, and on the other hand, tempo. More specifically, the fitted model coefficients indicate that, conditionally upon pitch interval deviation, the judges' ratings increase with lower F_0; moreover, the lower the pitch interval deviation, the larger the increase with lower F_0. Similarly, conditionally upon pitch interval deviation, the judges' ratings increase with faster tempo; the greater the pitch interval deviation, the larger the increase with faster tempo. As Table 2 reveals, tempo is involved in many significant pairwise interaction terms, which is a sign that it plays a central role in judges' ratings. Moreover, these interactions confirm that there are no direct relationships between the performance and quality parameters and the judges' ratings.

Concerning the effect of judge, it is noticeable that, although this factor interacts with several other covariates, the significant interaction terms relate to only a few judges (at most three of them). This means that for most judges there is absolutely no interaction effect; in other words, scoring does not vary across all judges. Rather, a few judges tend to exhibit significant interaction effects with one or more other covariates (F_0, tempo, VR and VE), meaning that, for these judges, some quality parameters are most significant for the scoring process. Note that there was no interaction between pitch interval deviation and judge. Indeed, the overall effect of pitch interval deviation on the judges' ratings did not vary across judges. This result is in line with the inter-judge reliability analysis in terms of pitch accuracy evaluation.

Discussion

This study supports the hypothesis that music experts are able to rate the pitch accuracy of operatic voices. Indeed, by observing the intra-judge reliability, we found that only two judges were not consistent between the test and the retest. In other words, each music expert seems to use similar strategies to rate the overall vocal pitch accuracy of the music material at different times. Future research on operatic singing voices could therefore be limited to one task instead of two. Note that the two inconsistent judges (judges 2 and 12 in Figure 3) did not show any particular features concerning their biographical data and music background or training compared to the other participants. Thus, we hypothesized that they had changed their rating strategy between the test and the retest and did not include these participants in the inter-judge analysis. The high inter-judge reliability does not confirm the variability pointed out by Sundberg et al. (1996). Note that this contrast could be explained by the nature of the task. Here we focused on the overall pitch accuracy evaluation of sung performances whereas Sundberg et al. (1996) examined the evaluation of individual tones. Our result suggests that the definition of global pitch accuracy was clear enough for all the music experts despite their different music backgrounds. The high intra- and inter-judge reliability could have been facilitated by the use of a pairwise paradigm. The efficiency of the pairwise comparison paradigm has been established as a means to evaluate disordered voices (Kacha, Grenez, & Schoentgen, 2005). Therefore its implementation seems appropriate for the perceptual judgments of operatic singing voices. Given that previous studies of such complex voices found it difficult to obtain a consensus among listeners (Ekholm et al., 1998; Garnier et al., 2007; Howes et al., 2004), this paradigm could be suggested as a means of achieving better reliability among the judges.

However, note that this experimental method involves the use of a limited number of short music excerpts to ensure that the perceptual task is a reasonable length.

These first analyses highlighted the music experts' ability to evaluate operatic voices and thus supported the possibility of computing a predictive model on the basis of the judges' ratings. Although it is visible in the model, the effect of the individual judge on the ratings is not great. Only a few judges exhibited a significant interaction with some of the acoustical parameters. This supports the hypothesis that some judges pay more attention or are more sensitive to certain acoustical parameters than others. This has an indirect effect on their ratings, but does so in a very limited way according to our analysis.

In the present study, the multiple regression modeling explains 78.8% of the variance in the judges' ratings. Compared to the study by Larrouy-Maestri et al. (2013) with untrained voices, the objective measurement of vocal pitch accuracy is not sufficient to predict the judges' rating in the case of operatic voices. Indeed, all covariates appeared at least once in the final model but the judges' rating cannot be explained by a single covariate, or by one covariate that does not interact with the others. For example, the pitch interval deviation parameter affects the judges' rating only when associated with the F_0 of the starting tone or with the tempo. Note that the music material in the present study (i.e., last sentence of "Happy Birthday") is not representative of the repertory of Western classical music. However, given that pitch interval deviation is not affected by the melodic effect (Larrouy-Maestri et al., in press), we can imagine that the performance of a melody from the Western classical repertory would lead to a similar pattern of results; it would be interesting to confirm this hypothesis.

In line with the conclusions of previous studies regarding the tolerance effect in intervallic and individual tones contexts (Sundberg et al., 1996; Vurma & Ross, 2006), the present study confirms that sung performances could be "out of tune" but not perceived as such. In addition, our results show that the combination of several acoustical parameters influences the perceptual judgment of the sung performance. Note that our objective measurement of pitch accuracy was computed on the basis of an equal temperament. This standard remains a reference in Western tonal music and seems adapted to the evaluation of solo performances (Berkowska & Dalla Bella, 2009; Dalla Bella & Berkowska, 2009; Dalla Bella et al., 2007; Larrouy-Maestri & Morsomme, in press; Pfordresher & Brown, 2009; Pfordresher et al., 2010). However, future researches using different temperaments as referents would allow confirming its relevance in the case of operatic voices. In contrast with a previous study about occasional singers (Larrouy-Maestri et al., 2013), our objective measurement of vocal pitch accuracy cannot be applied without considering the performance and quality parameters of the operatic singing voice. These findings are directly applicable in pedagogical settings. For instance, when a singing student is considered to be "out of tune" in a melodic context, he or she could be advised to improve intonation but also to take vocal quality and interpretation into account in order to be perceived as "in tune." The observation of a larger sample of singers and music experts would make it possible to improve the predictive modeling and thus the pedagogical implications of this study. Note that the unexplained 21.2% of variance may be linked to other criteria such as rhythm accuracy or vocal perturbation (Butte, Zhang, Song, & Jiang, 2009), which could unintentionally intervene in the judgment of vocal pitch accuracy.

Concerning the perceptual process, several studies reported that spectral composition (Hutchins et al., 2012; Russo & Thompson, 2005; Vurma et al., 2010; Warrier & Zatorre, 2002) and vibrato (van Besouw et al., 2008) affect the perception of tones. The several interactions observed in our multiple regression modeling showed that the individual acoustical parameters cannot be observed separately. Thanks to the analysis of natural sung performances, this study observed music experts' perceptual judgment of pitch accuracy in a melodic context and pointed out the relevance of performance and quality parameters in the evaluation of operatic voices. However, further studies using different combinations of the performance and quality parameters in synthesized material would allow one to manipulate these parameters and thus to clarify the perception of pitch accuracy in the case of operatic singing voices.

Conclusion

This study focused on the evaluation of vocal pitch accuracy in operatic singing voices. By observing the perceptual judgment of 14 sung melodies performed by professional singers, this study highlights the high intra- and inter-judge reliability of music experts when evaluating the pitch accuracy of operatic voices. However, the results show that the perceptual ratings were not directly linked with the objective measures of pitch accuracy. Interestingly, several interactions between performance and quality parameters contribute to explain the judges' ratings. Indeed, both vocal quality (energy, vibrato rate and extent) and interpretation of the tune (tempo, F_0) contributed to the judgment of vocal pitch accuracy. In addition to the pedagogical applications of these findings, this study opens up certain research perspectives concerning the perception of complex stimuli such as operatic singing voices.

Acknowledgments

The authors thank the Centre Henri Pousseur in Liège, the Laboratories of Images, Signals and Telecommunication Devices (LIST) in Brussels, Guillaume Videlier for technical support, Peter Pfordresher for our constructive discussions, and Marion Nowak for her help with the data collection. We also would like to thank the action editor, Richard Ashley, and three anonymous reviewers for their insightful comments and suggestions for this article.

References

Barnes, J. J., Davis, P., Oates, J., & Chapman, J. (2004). The relationship between professional operatic soprano voice and high range spectral energy. *Journal of the Acoustical Society of America, 116*, 530–538.

Berkowska, M., & Dalla Bella, S. (2009). Reducing linguistic information enhances singing proficiency in occasional singers. *Annals of the New York Academy of Sciences, 1169*, 108–111.

Bretos, J., & Sundberg, J. (2003). Vibrato extent and intonation in professional Western lyric singing. *Journal of Voice, 17,* 343–352.

Butte, C. J., Zhang, Y., Song, H., & Jiang, J. J. (2009). Perturbation and nonlinear dynamic analysis of different singing styles. *Journal of Voice, 23*, 647–652.

Dalla Bella, S., & Berkowska, M. (2009). Singing proficiency in the majority. *Annals of the New York Academy of Sciences, 1169*, 99–107.

Dalla Bella, S., Giguère, J.-F., & Peretz, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America, 121*, 1182–1189.

de Almeida Bezerra, A., Cukier-Blaj, S., Duprat, A., Camargo, Z., & Granato, L. (2009). The characterization of the vibrato in lyric and sertanejo singing styles: Acoustic and perceptual auditory aspects. *Journal of Voice, 23*, 666–670.

Ekholm, E., Papagiannis, G. C., & Chagnon, F. P. (1998). Relating objective measurements to expert evaluation of voice quality in western classical singing: Critical perceptual parameters. *Journal of Voice, 12*, 182–196.

Garnier, M., Henrich, N., Castellengo, M., Sotiropoulos, D., & Dubois, D. (2007). Characterisation of voice quality in Western lyrical singing: From teachers' judgements to acoustic descriptions. *Journal of Interdisciplinary Music Studies, 1*, 62–91.

Hirano, M., Hibi, S., and Hagino, S. (1995). Physiological aspects of vibrato. In P. Dejonckere, M. Hirano, & J. Sundberg (Eds.), *Vibrato* (pp. 9–34). San Diego, CA: Singular Publishing Group.

Howard, D. M. (2007a). Equal or non-equal temperament in a capella SATB singing. *Logopedics, Phoniatrics, Vocology, 32*, 87–94.

Howard, D. M. (2007b). Intonation drift in a capella soprano, alto, tenor, bass quartet singing with key modulation. *Journal of Voice, 21*, 300–315.

Howes, P., Callaghan, J., Davis, P., Kenny, D., & Thorpe, W. (2004). The relationship between measured vibrato characteristics and perception in Western operatic singing. *Journal of Voice, 18*, 216–230.

Hutchins, S., Roquet, C., & Peretz, I. (2012). The vocal generosity effect: How bad can your singing be? *Music Perception, 30*, 147–159.

Kacha, A., Grenez, F., & Schoentgen, J. (2005). Voice quality assessment by means of comparative judgments of speech tokens. *Proceedings of Interspeech 2005: 9th European Conference on Speech Communication and Technology*, 1733–1736.

Larrouy-Maestri, P., Lévêque, Y., Schön, D., Giovanni, A., & Morsomme, D. (2013). The evaluation of singing voice accuracy: A comparison between subjective and objective methods. *Journal of Voice, 27*, 259e1-259e5.

Larrouy-Maestri, P., Magis, D., & Morsomme, D. (in press). The effect of melody and technique on the singing voice accuracy of trained singers. *Logopedics, Phoniatrics, Vocology.*

Larrouy-Maestri, P., & Morsomme, D. (in press). Criteria and tools for objectively analysing the vocal accuracy of a popular song. *Logopedics, Phoniatrics, Vocology.*

Omori, K., Kacker, A., Carroll, L. M., Riley, W. D., & Blaugrund, S. M. (1996). Singing power ratio: Quantitative evaluation of singing voice quality. *Journal of Voice, 10*, 228–235.

Pfordresher, P. Q., & Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception and Psychophysics, 71*, 1385–1398.

Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M., & Liotti, M. (2010). Imprecise singing is widespread. *Journal of the Acoustical Society of America, 128*, 2182–2190.

Prame, E. (1994). Measurements of the vibrato rate of ten singers, *Journal of Acoustical Society of America, 94,* 1979–1984.

Prame, E. (1997). Vibrato extend and intonation in professional Western lyric singing. *Journal of Acoustical Society of America, 102*, 616–621.

Russo, F. A., & Thompson, W. F. (2005). An interval size illusion: The influence of timbre on the perceived size of melodic intervals. *Perception and Psychophysics, 67*, 559–568.

Seashore, C. E. (1938). *Psychology of music.* New York: McGraw Hill.

Stone, R. E., Cleveland, T. F., Sundberg, J., & Prokop, J. (2002). Aerodynamic and acoustical measures of speech, operatic, and Broadway vocal styles in a professional female singer. *TMH-QPSR, 43*, 17–29.

Sundberg, J. (1994). Acoustic and psychoacoustic aspects of vocal vibrato. *STL-QPSR, 35*, 45–68.

Sundberg, J. (1995). The singer's formant revisited. *STL-QPSR, 36*, 83–96.

Sundberg, J. (2001). Level and center frequency of the singer's formant. *Journal of Voice, 15,* 176–186.

Sundberg, J. (2013). Perception of singing. In D. Deutsch (Ed.), *The psychology of music* (pp. 69–105). San Diego, CA: Academic Press.

Sundberg, J., Prame, E., & Iwarsson, J. (1996). Replicability and accuracy of pitch patterns in professional singers. In P. J. Davis & N. H. Fletcher (Eds.), *Vocal fold physiology, controlling complexity and chaos* (pp. 291–306). San Diego, CA: Singular Publishing Group.

Thorpe, C. W., Cala, S. J., Chapman, J., & Davis, P. J. (2001). Patterns of breath support in projection of the singing voice. *Journal of Voice, 15*, 86–104.

van Besouw, R. M., Brereton, J. S., & Howard, D. M. (2008). Range of tuning for tones with and without vibrato. *Music Perception, 26*, 145–155.

Vurma, A., Raju, M., & Kuuda, A. (2010). Does timbre affect pitch? Estimations by musicians and non-musicians. *Psychology of Music, 39*, 291–306.

Vurma, A., & Ross, J. (2006). Production and perception of musical intervals. *Music Perception, 23*, 331–334.

Warrier, C. M., & Zatorre, R. J. (2002). Influence of tonal context and timbral variation on perception of pitch. *Perception and Psychophysics, 64*, 198–207.

Watts, C., Barnes-Burroughs, K., Andrianopoulos, M., & Carr, M. (2003). Potential factors related to untrained singing talent: A survey of singing pedagogues. *Journal of Voice, 17*, 298–307.

Author Note

Correspondence concerning this article should be addressed to Pauline Larrouy-Maestri, Département de Psychologie: cognition et comportement, Université de Liège, B-38, Rue de l'Aunaie, 30, 4000 Liège, Belgium. Electronic mail may be sent to pauline.larrouy@ulg.ac.be.

Table 1

Values of Each Singer's Objectively Measured Performance and Quality Parameters

	Performance parameters			**Quality parameters**		
Singers	Pitch interval deviation (in cents)	Tempo (bpm)	F_0 (in Hertz)	Energy distribution (ratio)	Vibrato rate (in Hertz)	Vibrato extent (in cents)
1	87.5	53.33	245.44	1.58	5.40	246
2	115.5	46.45	438.54	2.07	5.52	148
3	14.5	53.33	398.91	1.47	6.25	186
4	17	57.60	449.26	1.89	4.96	154
5	29	53.33	434.05	1.74	5.43	134
6	36.5	46.45	245.42	2.14	5.58	168
7	18	51.43	405.30	1.61	5.33	150
8	15.5	65.45	436.28	1.76	5.61	132
9	9.5	49.65	398.92	1.50	6.56	130
10	21	57.60	268.31	1.60	6.14	174
11	64	53.33	249.07	2.17	5.08	204
12	15	65.45	402.91	1.27	6.45	118
13	99.5	60.00	396.51	1.39	5.81	270
14	25.5	68.57	318.94	1.48	6.67	170

Table 2

Analysis of Variance of the Retained Multiple Linear Regression Model

Main effects	Sum Squares	DF	Mean Square	F-value	p-value
Pitch interval deviation	234.38	1	234.38	58.04	0.000
Judge	0.00	21	0.00	0.00	1.000
Tempo	20.32	1	20.32	5.03	0.030
F0	89.86	1	89.86	22.25	0.000
Energy	93.34	1	93.34	23.12	0.000
VR	16.13	1	16.13	3.99	0.050
VE	39.98	1	39.98	9.90	0.000
Pairwise interactions	Sum Squares	DF	Mean Square	F-value	p-value
Pitch interval deviation x F0	46.60	1	46.60	11.54	0.000
Pitch interval deviation x Tempo	607.97	1	607.97	150.56	0.000
Judge x Tempo	141.08	21	6.72	1.66	0.040
Judge x F0	82.27	21	3.92	0.97	0.500
Judge x VR	120.98	21	5.76	1.43	0.110
Judge x VE	267.36	21	12.73	3.15	0.000
Tempo x F0	2.94	1	2.94	0.73	0.390
Tempo x Energy	33.11	1	33.11	8.20	0.000
Tempo x VR	175.86	1	175.86	43.55	0.000
Tempo x VE	666.12	1	666.12	164.97	0.000
Residuals	767.21	190	4.04	///	///

Figures Caption

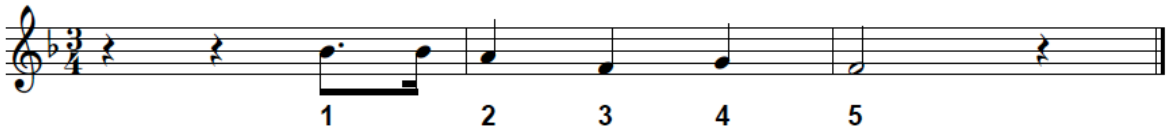

Figure 1. Music score representing the last sentence of the tune “Happy Birthday” with the number of the tones used to calculate the pitch interval deviation parameter.

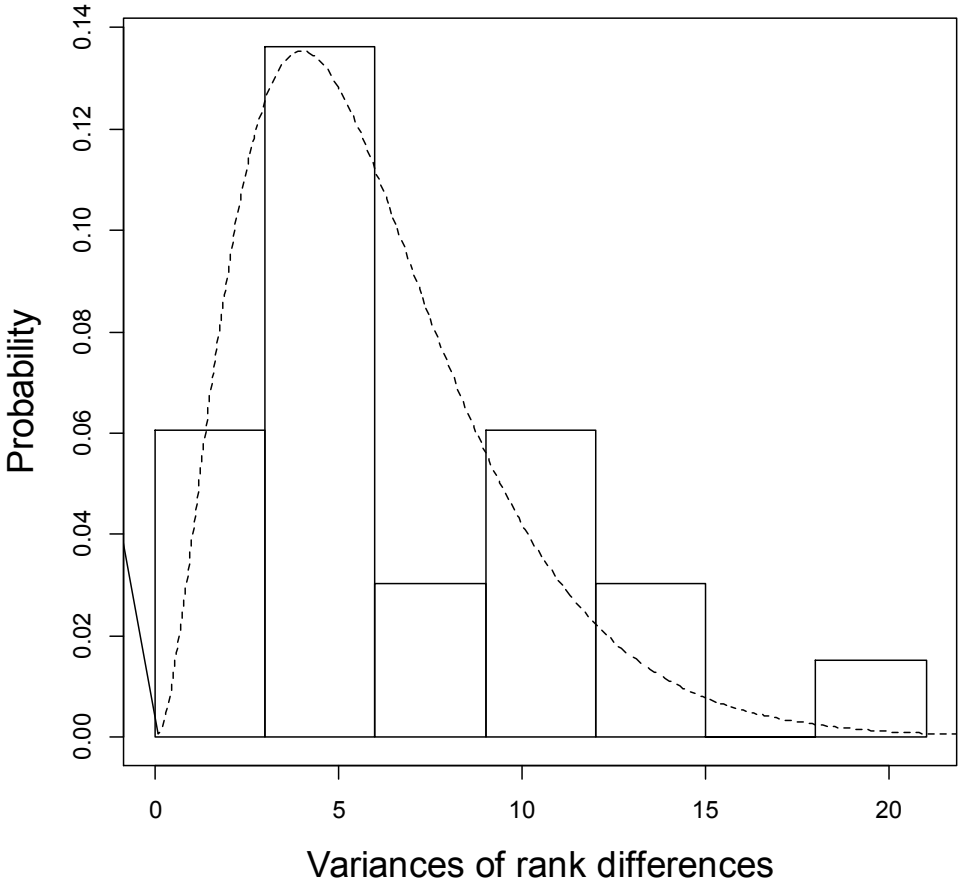

Figure 2. Sample distribution of the variances of rank differences. The dashed line represents the probability of the chi-squared distribution with six degrees of freedom.

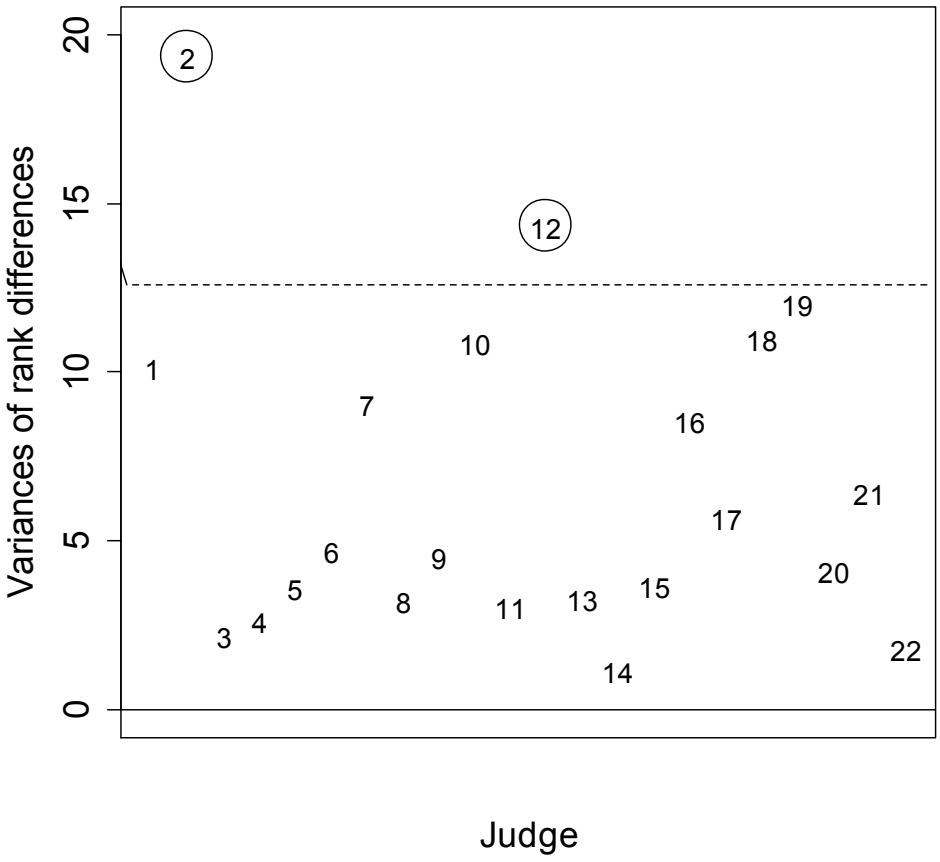

Figure 3. Variances of rank differences by judge, displayed by their number. The dashed line represents the cutoff value for identifying intra-judge reliability issues.

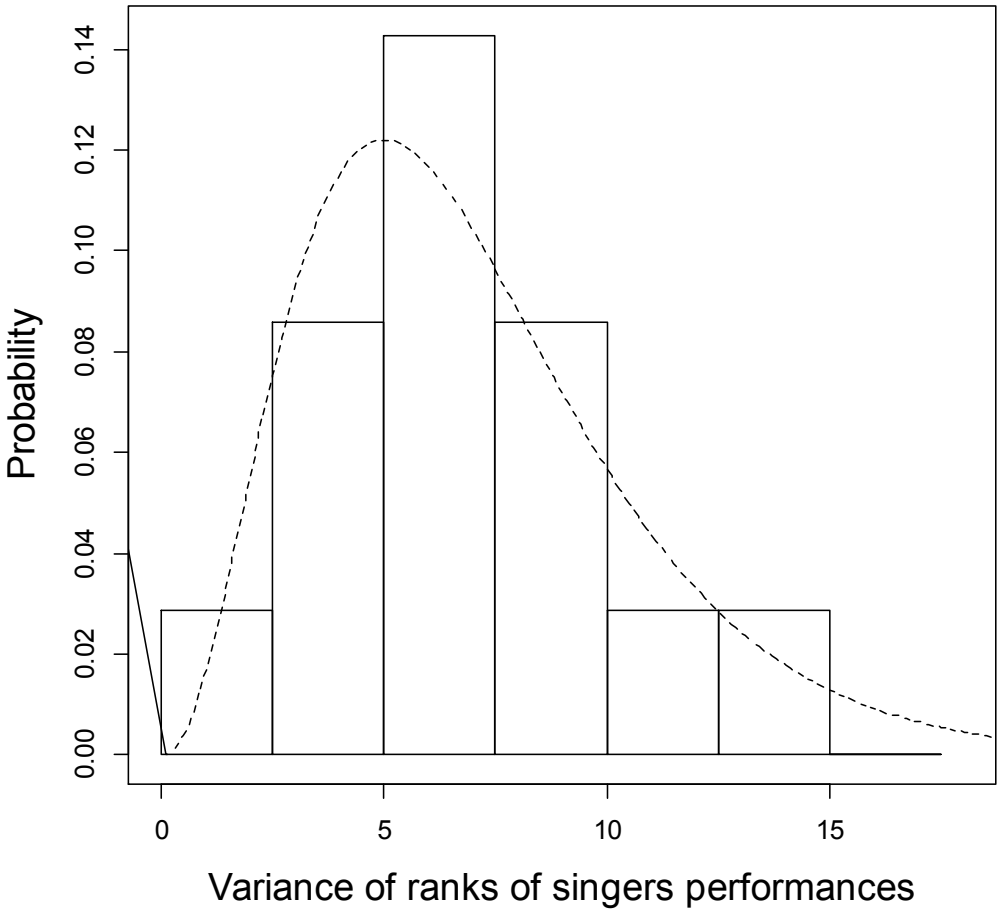

Figure 4. Sample distribution of the variances of rank of the sung performances. The dashed line represents the probability of the chi-squared distribution with seven degrees of freedom.

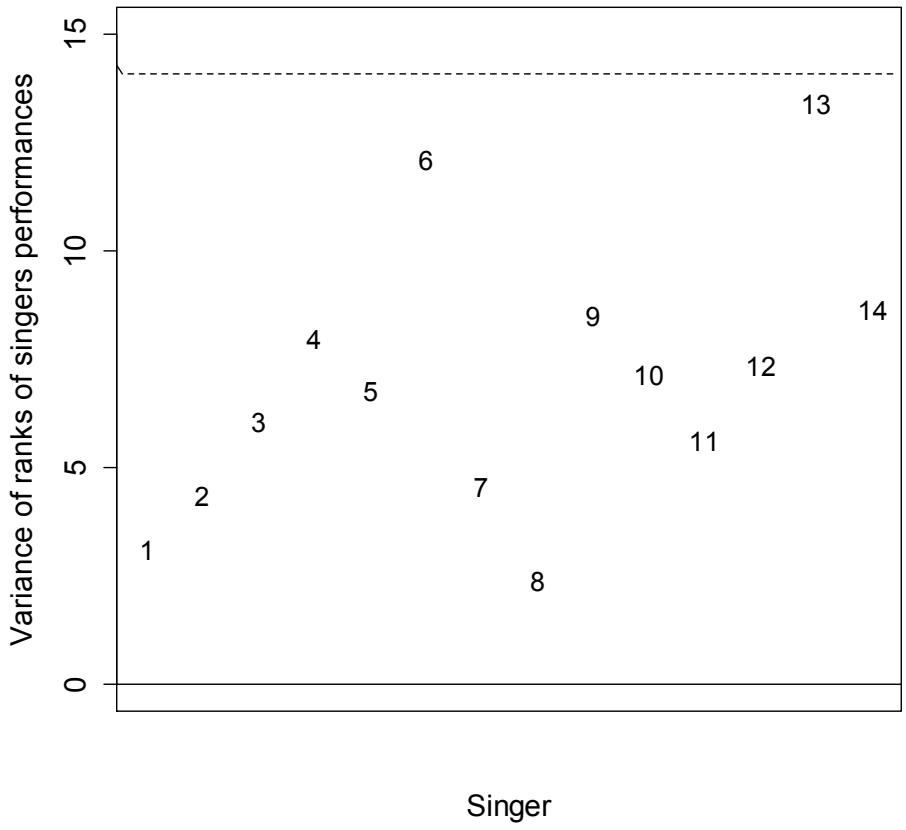

Figure 5. Variances of the rankings of sung performances, displayed by their number. The dashed line represents the cutoff value for identifying inter-judge reliability issues at the singer level.