LAY LISTENERS CAN EVALUATE THE PITCH ACCURACY OF OPERATIC VOICES

PAULINE LARROUY-MAESTRI
Max-Planck-Institute for Empirical Aesthetics, Frankfurt, Germany

DOMINIQUE MORSOMME
University of Liège, Liège, Belgium

DAVID MAGIS
Fonds de la Recherche Scientifique-FNRS, Liège, Belgium

DAVID POEPPEL
Max-Planck-Institute for Empirical Aesthetics, Frankfurt, Germany & New York University

LAY LISTENERS ARE RELIABLE JUDGES WHEN evaluating pitch accuracy of occasional singers, suggesting that enculturation and laypersons’ perceptual abilities are sufficient to judge “simple” music material adequately. However, the definition of pitch accuracy in operatic performances is much more complex than in melodies performed by occasional singers. Furthermore, because listening to operatic performances is not a common activity, laypersons’ experience with this complicated acoustic signal is more limited. To address the question of music expertise in evaluating operatic singing voices, listeners without music training were compared with the music experts examined in a recent study (Larrouy-Maestri, Magis, & Morsomme, 2014a) and their ratings were modeled with regard to underlying acoustic variables of pitch accuracy. As expected, some participants lacked test-retest reliability in their judgments. However, listeners who used a consistent strategy relied on a definition of pitch accuracy that appears to overlap with the quantitative criteria used by music experts. Besides clarifying the role of music expertise in the evaluation of melodies, our findings show robust perceptual abilities in laypersons when listening to complex signals such as operatic performances.

Received: November 24, 2015, accepted October 17, 2016.

Key words: music expertise, pitch accuracy, melodic perception, singing voice, opera

MELODIES PERFORMED WITH WESTERN operatic technique yield a complex acoustic signal (see Larrouy-Maestri, Magis, & Morsomme, 2014b; Sundberg, 2013, for reviews). Therefore it is unsurprising that the perception of pitch accuracy is not solely related to the information outlined in a musical score (i.e., the pitch relationship between tones), which is the main musical criterion used to estimate singing proficiency in occasional singers (see Dalla Bella, 2015, or Pfordresher & Larrouy-Maestri, 2015, for discussions on this topic). Operatic performances also include several acoustic parameters, such as vibrato or singer’s formant, developed through operatic vocal training. In order to examine the construct of pitch accuracy in operatic voices, Larrouy-Maestri et al. (2014a) asked professional musicians to rate the pitch accuracy of “Happy Birthday” performances sung with an operatic singing technique. This study showed that pitch accuracy evaluation was not based exclusively on the precision of performed music intervals but on a complex combination of performance variables (i.e., tempo, fundamental frequency of the starting tone) and quality parameters (i.e., energy distribution, vibrato rate, and vibrato extent). For instance, large pitch interval deviations influenced the judges’ rating only when associated with the pitch height of the starting tone or with the tempo. The results confirmed the complexity of singing accuracy perception but also highlighted the high intra and interjudge reliability and the “objectivity” of the listeners (i.e., judgments relying mainly on the quantifiable variables mentioned above). Interestingly, these qualities (i.e., objectivity and consistency) are not exclusive to professionals with intense formal training. Recently, Larrouy-Maestri, Magis, Grabenhorst, and Morsomme (2015) observed that lay listeners behaved similarly when evaluating the pitch accuracy of melodies performed by occasional singers. This finding demonstrates perceptual abilities in lay listeners comparable to musically trained listeners and appropriate use of implicitly learned musical rules (e.g., Bigand & Poulincharronnat, 2006; Hannon & Trainor, 2007; Marmel, Tillmann, & Dowling, 2008). However, examining the evaluation of occasional singers might not
suffice to conclude that lay listeners are objective (i.e., relying on quantifiable definitions of pitch accuracy) and reliable (i.e., consistent). Indeed, more challenging conditions could lead to different results (Besson & Faita, 1995) and therefore reflect the limits of either perceptual abilities or use of internalized musical rules when listening to complex auditory signals.

Evaluating operatic performances is challenging for two main reasons. First, opera is not a popular music style. According to the National Endowment for the Arts (2008), only 2.1% of the adult population of the United States attend an opera performance per year. Although there are operatic voice contests (e.g., Queen Elisabeth International Music Competition of Belgium), such events are rare compared to the numerous casting shows and talent contests in the media. Therefore, one can infer that laypersons are not particularly familiar with this specific style and its acoustic features. For instance, the singer’s formant consists of increasing the energy between 2 and 4 kHz (Barnes, Davis, Oates, Chapman, 2004; Omori, Kacker, Carroll, Riley, & Blaugrund, 1996; Thorpe, Cala, Chapman, & Davis, 2001) and is quite typical for this vocal technique (Larrouy-Maestri et al., 2014b; Stone, Cleveland, Sundberg, & Prokop, 2002). Developed during training (Brown, Rothman, & Sapienza, 2000; Lundy, Roy, Casiano, Xue, & Evans, 2000; Omori et al., 1996), this feature is considered a “vocal quality” by singing teachers (Ekholm, Papagiannis, & Chagnon, 1998; Garnier, Henrich, Castellengo, Sotropoulos, & Dubois, 2007) and allows singers to be heard over an orchestra without electronic amplification. Since most popular music styles make use of amplification, this acoustic parameter is expected to be quite unfamiliar to lay listeners. Note that other parameters such as vibrato—also considered a “vocal quality” developed through training (Mürbe, Zahnert, Kuhlisch, & Sundberg, 2007)—are not exclusive to opera but occur in other singing styles too. Lay listeners might thus be acquainted with this feature even if they are not opera listeners. Second, the definition of pitch accuracy in operatic singing voices is based on the interaction of several parameters (Larrouy-Maestri et al., 2014a), in contrast with the definition of pitch accuracy of melodies performed by occasional singers (Larrouy-Maestri, Lévêque, Schön, Giovanni, & Morsonme, 2013; Larrouy-Maestri et al., 2015). In addition, parameters such as the spectral composition of the sound (e.g., Hutchins, Roquet, & Peretz, 2012; Russo & Thompson, 2005; Vurma, Raju, Kuuda, 2010; Warrier & Zatorre, 2002) or vibrato (van Besouw, Brereton, & Howard, 2008) influence pitch perception. The combination of these parameters in a single signal may enhance the complexity of the task for laypersons and challenge their perceptual abilities, which are typically deemed weaker than music experts’ (e.g., Kraus & Chandrasekaran, 2010; Michely, Delhomméau, Perrot, & Oxenham, 2006; Schellenberg & Weiss, 2013; Tervaniemi, Just, Koelsch, Widmann, & Schroger, 2005; Varnet, Wang, Peter, Meunier, & Hoen, 2015; but see Larrouy-Maestri et al., 2015; Vanden Bosch der Nederlanden, Hannon, & Snyder, 2015, for contradictory findings).

To investigate whether music expertise is a prerequisite for objective and consistent evaluation of pitch accuracy in challenging conditions, we asked lay listeners to evaluate, in a test-retest paradigm, the pitch accuracy of operatic singers previously analyzed in Larrouy-Maestri et al. (2014a); we then modeled their performance with regard to the acoustic predictors of pitch accuracy definition and to the music experts data of the reference study (Larrouy-Maestri et al., 2014a).

Method

PARTICIPANTS

Twenty-two participants were paired by gender (8 women) and age ($M = 45.59$ years, $SD = 11.64$) with the expert listeners of the reference study (Larrouy-Maestri et al., 2014a). They were recruited in Belgium and France on the basis of an extended questionnaire. None mentioned possessing absolute pitch, they reported no history of choral singing, no history of formal music training (or a maximum of 2 years of training and no practice during the past 5 years) and no particular affinity for music (attending less than one concert a week and actively listening to music less than two hours a day). We applied additional inclusion criteria before enrolling participants in the present study: (a) bilateral hearing threshold of 20 dB SPL at 500, 1000, 2000, and 4000 Hz, screened with pure tone audiometry (Madsen Xeta, GN Otometrics, Denmark), (b) no congenital amusia (tested with the Montreal Battery of Evaluation of Amusia of Peretz, Champod, & Hyde, 2003), and (c) the ability to perform the song “Happy Birthday” with the appropriate melodic contour.

MATERIAL

The material consisted of 14 performances from the corpus http://sldr.org/sldr000792/en (last sentence of a cappella “Happy Birthday” performed with an operatic singing technique by 14 female professional singers, from 21 to 66 years old, $M = 34.86$, $SD = 12.72$). The performers reported on average 10.57 years of formal vocal training ($SD = 3.58$), regular singing practice, and
public solo performances in classical style at the time of the recording. As summarized in the table, each performance was analyzed with respect to pitch interval deviation (mean absolute value of the deviation between interval performed and expected, in cents), tempo (in beats per minute), F0 of the starting tone (in Hertz), energy distribution (ratio between the 2.4-5.4 kHz band and the 0-2.4 kHz band, associated with the presence of singers' formant), vibrato rate (number of quasiperiodic modulations of the F0 per second, in hertz) and extent (amplitude of the F0 variations within the same tone, in cents). Note that the pitch interval deviation measurement provides an adequate representation of deviations (i.e., great variability in Table 1) over short melodies (i.e., not long enough to measure the precision over time or the respect of the tonal center). Individual values for each performance can be found in the reference paper (Larrouy-Maestri et al., 2014a).

A test-retest procedure was applied after 8 to 15 days (M = 9.32 days).

ANALYSIS
To examine intrajudge reliability, the 14 performances were ranked from worst to best, separately by each nonexpert judge, at two timepoints (test and retest). Then, rank differences (test minus retest) were computed for each performance, and the sample variance of differences was computed for each judge. This yields 22 sample variances, each variance being computed on 14 rank differences. Large variances indicate large dispersion in rank differences, and thus large variability in the rankings on test and retest and consequently low intrajudge reliability. Cut-off values for sampling variances in reliability analyzes were similar to those derived from expert judges of the reference study (Larrouy-Maestri et al., 2014a): Variances were compared to a chi-square 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 .95. A similar process was used to measure interjudge reliability, but using only data of the first test, disregarding the retest. Sample variances were computed for each singer on the set of rankings derived from each nonexpert judge. Low variance indicates low variability in the judges’ rankings of the singers, and thus good interjudge reliability.

In order to compare lay listeners’ and professional musicians’ definition of pitch accuracy, expert and nonexpert judges’ ratings were modeled by combining the currently analyzed data from nonexpert judges with the data collected in the reference paper, Larrouy-Maestri et al. (2014a). In this multiple linear model with judges’ ratings as independent variables, all acoustic factors (pitch interval deviation, tempo, F0 of the starting tone,

<table>
<thead>
<tr>
<th>Acoustic parameters</th>
<th>Mean (SD)</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch interval deviation (cents)</td>
<td>40.57 (35.72)</td>
<td>9.50</td>
<td>115.50</td>
</tr>
<tr>
<td>Tempo (bpm)</td>
<td>55.86 (6.96)</td>
<td>46.45</td>
<td>68.57</td>
</tr>
<tr>
<td>F0 (Hertz)</td>
<td>363.42 (79.51)</td>
<td>245.42</td>
<td>449.26</td>
</tr>
<tr>
<td>Energy distribution (ratio)</td>
<td>1.69 (0.28)</td>
<td>1.39</td>
<td>2.17</td>
</tr>
<tr>
<td>Vibrato rate (Hertz)</td>
<td>5.77 (0.55)</td>
<td>4.96</td>
<td>6.67</td>
</tr>
<tr>
<td>Vibrato extent (cents)</td>
<td>170.29 (44.19)</td>
<td>118.00</td>
<td>270.00</td>
</tr>
</tbody>
</table>

Note: All the acoustical data analyses were performed on a Macintosh (Mac OS X, Version 10.6.8) with AudioSculpt 2.9.4v3 and OpenMusic 6.3 softwares (IRCAM, Paris, France) using a Short Time Fourier Transform (STFT) analysis. The pitch interval deviation measurement was computed with regard to equal temperament and reported as a continuous variable. Absolute values of the deviations between the non-repeated tones of the short melodies were preferred to signed values in order to avoid the cancellation of enlarged/compressed intervals and therefore the underrepresentation of pitch deviations in trained singers. Note that energy distribution, vibrato rate and extent were quantified on the basis of the longest tone (i.e., last tone) of the last sentence of “Happy Birthday” (length from 1.13 to 1.98 seconds, M = 1.45, SD = 0.09) in order to get enough acoustic information for adequate measurements.
energy distribution, vibrato rate and vibrato extent) as well as a group effect (with categories “expert” and “nonexpert”) were included as covariates. All possible main effects and pairwise interactions were set up in the complete model, while the effect of a single judge was not modeled. The complete model was simplified by removing all nonsignificant terms, leading to the simplest model retained for analysis. Significant effects and their direction were assessed by Wald tests with appropriate contrast matrices across model parameters and limiting chi-square distributions (Rao, 1973). Note that in this analysis the Wald statistics are referred to as Q statistics, due to the use of contrast matrices.

**Results and Discussion**

### INTRA AND INTERJUDGE RELIABILITY IN LAY LISTENERS

Six of the 22 nonexpert participants exhibited high variability in their rank differences and exceeded the expert-based threshold (Figure 1A). This indicates weaker intrajudge reliability in the group of nonexpert judges. Note that the music experts in Larrouy-Maestri et al. (2014a) did also not show “perfect” intrajudge reliability (e.g., two of them seemed to change their rating strategy between test and retest). However, judges who use a similar strategy on both test and retest evidently share a similar strategy with the other judges. Indeed, Figure 1B shows that all sample variances are below the limit threshold derived from the reference study, indicating good overall interjudge reliability in the nonexpert group. Thus, most of laypersons are capable to develop a common “construct” of pitch accuracy despite the complexity of the signal and despite their limited experience as listeners to this specific style of singing.

**LAY LISTENERS’ VERSUS PROFESSIONAL MUSICIANS’ DEFINITION OF PITCH ACCURACY**

The final model calculated to explain the judges’ ratings has an $R^2$ coefficient of .53, very close to the coefficient of the full model (.532), meaning that the acoustic variables under study explained more than half of the variance of the ratings. The final model (illustrated in Figure 2) can be interpreted as follows. First, all covariates are included in the model, but none of them as simple main effect. Judges’ ratings cannot be fully described with individual covariates but with a complex interaction structure instead. Second, several covariates partly explain the ratings as pairwise interactions, though such interactions are independent of the particular group of judges (experts or nonexperts). All those pairwise interactions involve the tempo on one hand, and on the other hand either pitch interval deviation, $Q(1) = 7.87, p = .005$, F0 of the starting tone, $Q(1) = 120.96, p < .001$, vibrato rate, $Q(1) = 135.31, p < .001$, or vibrato extent, $Q(1) = 41.49, p < .001$. Note that the limited contribution of the pitch interval deviation parameter in the model could not be attributed to low variability among singers (see Table 1).
Directional interpretations of these pairwise interactions are as follows (increase/decrease is highlighted in Figure 2 with positive and negative symbols): At a given level of pitch interval deviation, ratings will increase with a decrease of tempo; while at a given level of either F0 of the starting tone, vibrato rate or vibrato extent, ratings will increase with an increase of tempo. The overall results highlight the complexity of the definition for experts and nonexperts as well as the large overlap between the two groups, as illustrated by the solid lines in Figure 2. In addition to the relative reliability pointed out in the previous analysis, lay listeners seem to use criteria that are also important to the music experts, despite the complexity of the signal to be evaluated and the unfamiliar character of operatic performances for these participants.

Finally, several pairwise interactions of covariates have different effects according to the respective group of judges, as shown by the dashed lines in Figure 2. These pairwise interactions involve: Tempo and energy distribution, Q(1) = 25.47, p < .001, pitch interval deviation and F0 of the starting tone, Q(1) = 9.93, p = .003, and pitch interval deviation and energy, Q(1) = 31.94, p < .001. Concretely, at given levels of energy, judges’ ratings increase with tempo, both for nonexpert judges, Q(1) = 77.48, p < .001, and for expert judges, Q(1) = 144.20, p < .001, but the increase is more pronounced for experts than for nonexperts. Also, at given levels of pitch interval deviation, ratings increase with F0 of the starting tone in expert judges, Q(1) = 9.42, p = .002, but remain constant in nonexpert judges, Q(1) = 0.54, p = .46. In addition, at a given level of pitch interval deviation, ratings increase with a decrease in the energy distribution measure for nonexpert judges, Q(1) = 27.69, p = .002, but remain constant for expert judges, Q(1) = 2.95, p = .09.

It is tempting to attribute the differences occurring between the two groups of judges to weaker perceptual abilities in nonmusicians. For instance, the fact that lay listeners give better ratings to lower energy distribution for an equal pitch interval deviation value could be explained by a greater influence of timbre (partly explained by the energy distribution in a signal) on pitch perception (e.g., Hutchins et al., 2012; Russo & Thompson, 2005; Vurma et al., 2010; Warrier & Zatorre, 2002). However, lower pitch perception abilities would also be noticeable due to an expertise effect on other interactions including vibrato rate or extent or tempo, which are also known to affect pitch perception (Prince, 2011; van Besouw et al., 2008). The model here suggests that interactions including these parameters do not differ between the two groups, which weakens the explanation of expertise effects by lower perceptual abilities in laymen, at least as far as listening to operatic voices is concerned. Further studies with materials allowing multiple parameters and their interactions to be controlled are necessary to fully eliminate this explanation. Nevertheless, the present findings already support that the differences observed in our model could be interpreted as subtle differences in the definition of pitch accuracy, as a consequence of specific knowledge of operatic singing performances. Indeed, specific knowledge regarding the technical difficulties of operatic performances could explain the relevance of the interactions “pitch interval deviation - F0 of the starting tones” (better ratings when the F0 is high, for similar pitch interval deviation) and “energy – tempo” (better ratings when the tempo is fast, for similar energy distribution) for experts. Singer’s formant (estimated through the energy measurement) occurs on vocalic sounds and, as fast tempo shortens vowels, music experts would acknowledge the difficulty
of singing faster while maintaining energy distribution in the signal. As a consequence, they would grant the performers higher ratings. A similar explanation seems plausible regarding the interaction “pitch interval deviation - F0 of the starting tones.” Indeed, the development of the vocal range is a key aspect in vocal training, and starting the melody on a high pitch could be seen as a special vocal effort that requires both breath support and control of laryngeal muscles. Here again, specific knowledge about vocal technique itself could lead music experts to incorporate the vocal effort required to sing in a high range for similar precision of the pitch intervals. This would explain why they rated these performances better than lay listeners did. Moreover, less knowledge of lay persons could explain the interaction “pitch interval deviation - energy distribution.” Since lay persons have limited exposure to operatic voices and knowledge about the performance settings (i.e., no amplification), they would show preference for what they simply know and interpretation of low energy distribution as more adequate for similar interval precision in the signal. Interestingly, our model did not show differences between music experts and nonexperts regarding the contribution of the vibrato rate and vibrato extent parameters. It might be the case that the effect of formal music expertise on music perception is exclusively limited to acoustic parameters that lay listeners simply do not know. The differences between lay and professional listeners seem more attributable to the accrual of domain knowledge about vocal technique than to differences in perceptual abilities or different definitions of pitch accuracy.

**Conclusion**

We examined the ability of listeners to evaluate pitch accuracy of complex signals. We applied the approach described in Larrouy-Maestri et al. (2014a) to matched nonexperts. The results demonstrate surprising consistency and “objectivity” of the listeners despite the complexity of the signal evaluated. Whereas lay listeners make trustworthy judges when evaluating melodies performed by their peers (Larrouy-Maestri et al., 2015), here we observed two different profiles among the lay listeners: listeners who cannot maintain a consistent strategy in evaluating complex voice versus listeners who behave similar to musicians (cf. Larrouy-Maestri et al., 2014a), although they lacked the technical expectations due to the absence of specific knowledge about this singing technique. Therefore, the role of formal music training seems limited when evaluating technical features (e.g., pitch accuracy in the present study). We suggest that using similar methods with a focus on aesthetic features might allow a better understanding of the role of formal music training in music appreciation.

**Author Note**

We thank the Centre Henri Pousseur in Liège, Laboratories of Images, Signals and Telecommunication Devices (LIST) in Brussels, and Felix Bernoully for technical support, Marion Nowak and Virginie Roig-Sanchis, for their help with the data collection, and Renan Marcello Vairo Nunes, Matthias Grabenhorst, Tina Roeske, and Wolf Schlotz for helpful comments on a previous version of the manuscript.

**Correspondence concerning this article should be addressed to Pauline Larrouy-Maestri, Neuroscience Department, Max-Planck-Institute for Empirical Aesthetics, 14 Grüneburgweg, 60322 Frankfurt am Main, Germany. E-mail: pauline.larrouy-maestri@aesthetics.mpg.de**

**References**


