



OPEN Perceptual (but not acoustic) features predict singing voice preferences

Camila Bruder¹✉, David Poeppel^{2,3,4} & Pauline Larrouy-Maestri^{1,4}

Why do we prefer some singers to others? We investigated how much singing voice preferences can be traced back to objective features of the stimuli. To do so, we asked participants to rate short excerpts of singing performances in terms of how much they liked them as well as in terms of 10 perceptual attributes (e.g.: pitch accuracy, tempo, breathiness). We modeled liking features based on these perceptual ratings, as well as based on acoustic features and low-level features derived from Music Information Retrieval (MIR). Mean liking ratings for each stimulus were highly correlated between Experiments 1 (online, US-based participants) and 2 (in the lab, German participants), suggesting a role for attributes of the stimuli in grounding average preferences. We show that acoustic and MIR features barely explain any variance in liking ratings; in contrast, perceptual features of the voices achieved around 43% of prediction. Inter-rater agreement in liking and perceptual ratings was low, indicating substantial (and unsurprising) individual differences in participants' preferences and perception of the stimuli. Our results indicate that singing voice preferences are not grounded in acoustic attributes of the voices per se, but in how these features are perceptually interpreted by listeners.

Singing is ubiquitous across cultures^{1,2} and highly popular³, but fundamental questions about our preferences remain unanswered. For instance: why do we like some singing voices more than others? How much are these preferences based on attributes of the singing voice, and how much are they related to listeners' internal factors?

The extent to which aesthetic appreciation depends on objective properties of a stimulus versus a person's internal subjective states and evaluations has long been debated. Pioneering empirical work has been done in the visual domain, starting with Fechner's investigations of average group preferences for rectangles following the golden ratio^{4,5}. More recently, some studies have argued for average preferences for certain stimulus features such as curved versus sharp contours⁶, symmetrical patterns⁷, and high contrast images⁸, among others—though variability across people tends to be large⁹. In fact, for certain types of stimuli (e.g., abstract paintings), individual differences are found to be particularly striking, with small interindividual agreement (or highly idiosyncratic taste) across participants^{10–12}, supporting the role of individuals' internal factors and their personal experiences as critical determinants of their preferences.

In the case of music, a broad range of stimulus features such as complexity, tempo, rhythm, tonal and timbral features, among others, have been shown to predict aesthetic responses to musical sounds (see¹³ for a review). Like in the visual domain, musical preferences are assumed best to be understood with an interactionist approach (e.g.:^{14,15}), that is, as the result of a complex interplay of numerous multilevel factors, which can be linked to attributes of the music stimuli themselves (i.e., acoustic features), as well as their perception by listeners (i.e., perceptual features) and other cognitive, emotional, evaluative processing on the listener side.

In the case of the human voice, empirical work has focused on preferences for the spoken voice: studies of vocal attractiveness indicate that the height of fundamental frequency (f_0) and formant dispersion of a given voice may function as indirect cues of body size, health, and age, with higher f_0 and more spread formants generally preferred for women's voices and lower f_0 preferred for men's voices¹⁶. A different line of research has used voice morphing to show that composite, averaged voices (which are smoother and have higher signal-to-noise ratio than individual voices) were considered more attractive than most of the individual voices presented to participants¹⁷; but see¹⁸.

Particularly for the singing voice, there is evidence that specific acoustic features influence the perception of pitch accuracy, or of singing performances as “correct”¹⁹ and that the use of such features is affected by the

¹Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany. ²New York University, New York, NY, USA. ³Ernst Strüngmann Institute for Neuroscience, Frankfurt, Germany. ⁴Max Planck-NYU Center for Language, Music, and Emotion (CLaME), New York, USA. ✉email: camila.bruder@ae.mpg.de

music expertise of the listeners^{20,21}—but such research does not answer questions about singing preferences. In fact, aesthetic appreciation probably goes beyond correctness, that is to say, people do not attend concerts to hear ‘correct’ performances, but to enjoy them. In line with findings about voice attractiveness, one might expect that appreciation of singing would depend heavily on the acoustic properties of voices/singing performances. On the other hand, literature on aesthetic responses to artistic performances (music as well as visual) supports idiosyncrasies on the listener’s side.

Here we aimed to quantify how much the aesthetic appreciation of singing voices depends on objective properties of the sung stimuli and to investigate which acoustic qualities—and their perceptual counterparts—are relevant in this process. To this end, we collected liking ratings (online and lab experiments) and perceptual ratings (lab experiment) of short excerpts of a cappella singing, and we examined, through statistical modeling, the relationship between liking ratings, acoustic properties, and perceptual ratings of these singing performances.

Results

Large variability in liking of singing performances

In a short online experiment (~ 12 min long), 326 participants rated on a 9-point scale how much they liked excerpts of two contrasting a cappella melodies (see Fig. 1A and “Methods” section for details), performed repeatedly by 16 highly trained female singers, for a total of 96 stimuli. Each stimulus was rated by 103, 106 or 117 participants. Despite the large variability in liking ratings, preferences for some stimuli emerged (Fig. 1C), with, e.g., low ratings for the performances of Singer 6 and higher ratings for the performances of Singer 16 (Fig. 1B).

To better understand the large variability of liking ratings observed in Fig. 1, we examined participants’ intra- and inter-rater reliability. The consistency of ratings in a subset of 16 repeated trials revealed a mean test–retest correlation of 0.41 ($SD=0.29$) and a distribution ranging from -0.61 to 0.95 . We computed Krippendorff’s alpha (α_K) for the (within-participant normalized) liking ratings and found low agreement across participants ($\alpha_K=0.10$, $\alpha_K=0.07$ based on raw ratings). Agreement was higher among 145 “consistent” participants (with $r_{\text{test-retest}}$ scores higher than 0.5; $\alpha_K=0.19$; $\alpha_K=0.14$ based on raw ratings). Note that this is in line with previous literature reporting variably low agreement between raters in perceptual tasks using music (e.g.,^{22,23} and voice stimuli²⁴). Please see Supplementary Figure S1 for the distribution of raw and within-participant normalized liking ratings.

Very limited role of acoustic features in singing voice preferences

To examine the role of acoustics in singing voice preferences, we constructed linear mixed models that included acoustic features (i.e., acoustic model) as well as features selected from Music Information Retrieval (MIR) and

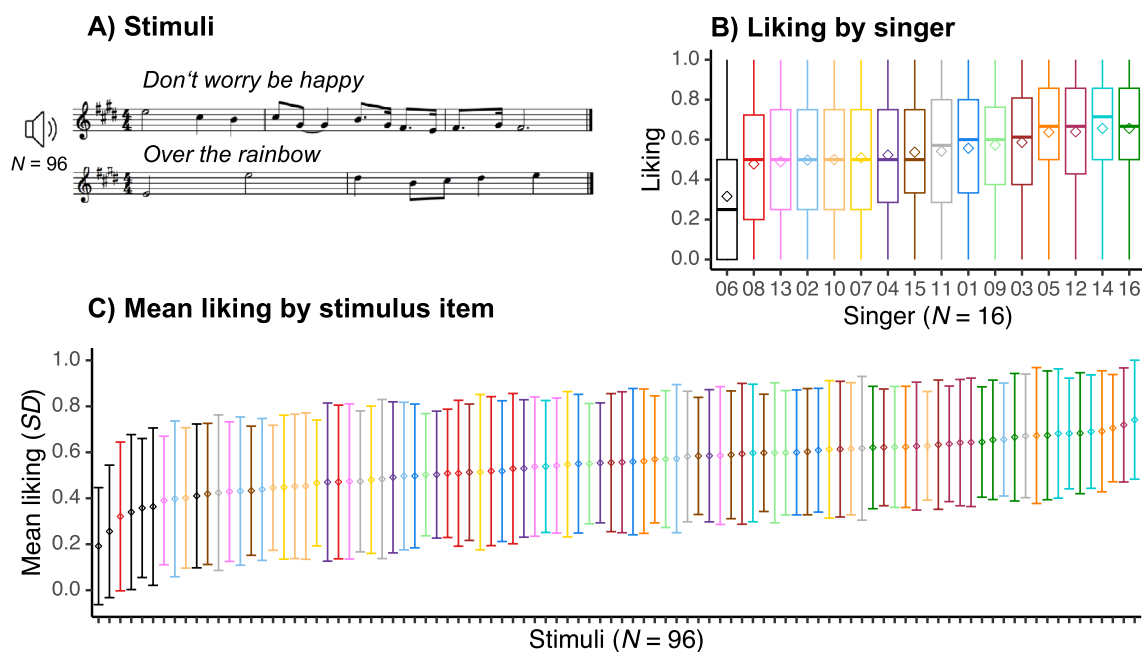


Figure 1. Online experiment. (A) Methods: 326 participants rated on a scale of 1 to 9 how much they liked short excerpts of two different melodies, performed at different times and without lyrics, by 16 trained female singers. (B) Boxplots of (within-participant normalized) liking ratings by singer. Colors correspond to individual singers. Diamonds depict average liking ratings. Lower and upper hinges correspond to the first and third quartiles, and whiskers extend from the hinge to $1.5 \times$ inter-quartile range (Tukey-style boxplot). (C) Mean (within-participant normalized) liking ratings ranked from least to most liked stimuli. Each stimulus was rated by 103, 106 or 117 participants. As in Fig. 1b, colors correspond to individual singers. Diamonds depict average liking ratings per stimulus item, and error bars correspond to one standard deviation above and below the average value.

from the Soundgen R package, which was developed to deal with human nonverbal vocalizations²⁵. See "Methods" section and Supplementary Information for details about MIR and Soundgen models, Supplementary Table S1 for descriptive statistics of acoustic features by melody and Supplementary Figure S1 for an illustration of the distribution of acoustic features by melody. The acoustic model predicting (within-participant normalized) liking ratings from pitch interval deviation, tempo, energy ratio, harmonics-to-noise ratio, vibrato rate, vibrato extent, jitter, shimmer, and cepstral peak prominence (CPP), and including random intercepts for participants as well as for stimulus items nested in singers, performed better than the null model ($\chi^2(9) = 23.597, p < 0.01$). However, the acoustic model predicted remarkably little variance in liking ratings (Supplementary Table S2), with almost all of the modeled variance captured by the random effects (marginal $R^2 = 0.016$, conditional $R^2 = 0.24$). Note that fitting the acoustic model to a subset of 145 "consistent" participants ($r_{\text{rest-retest}}$ scores > 0.5) led to similar results, that is, very low prediction based on fixed effects (marginal $R^2 = 0.025$, conditional $R^2 = 0.3$). This suggests that the low prediction was not simply a consequence of participants' inconsistent rating behavior, which could stem from inattention or lack of engagement to the task—common concerns in any experimental task, but especially so in the case of online experiments²⁶. When predicting liking ratings from selected features from MIR and from the Soundgen toolbox, we observed that the best performing MIR model and the Soundgen model also could not account for much of the variance in liking ratings (Supplementary Table S3: marginal $R^2 = 0.024$, conditional $R^2 = 0.227$ for the best performing MIR model; marginal $R^2 = 0.025$, conditional $R^2 = 0.223$ for the Soundgen model). Here again, fitting the same models on a subset of 145 consistent participants also led to low prediction based on fixed effects (marginal $R^2 = 0.047$, conditional $R^2 = 0.281$ for the MIR model and marginal $R^2 = 0.052$, conditional $R^2 = 0.273$ for the Soundgen model). In summary, the acoustic features under study have been shown to be directly relevant to listeners when evaluating the technical aspects of singing performances (i.e., correctness) and spoken voice attractiveness. However, the modeling approach revealed their very limited role in predicting variance of singing preferences.

Perceptual features predict variance in singing preferences

To explore how much aesthetic appreciation of singing is grounded in perceptual features of the voices—how listeners perceive acoustic characteristics—we collected ratings of the stimuli on several perceptual features in a lab experiment. Forty-two participants rated the same 96 stimuli on 10 different scales (Fig. 2 and see "Methods" section for details; see Supplementary Figure S3 for the distribution of normalized ratings). They also provided liking ratings, allowing us to replicate the results of the online experiment in a controlled setting. Similar to the online experiment, the distribution of liking ratings was spread and a gradual pattern of average preferences emerged (Supplementary Figure S4b and c). Mean liking ratings per stimulus item correlated highly with the ones from the online experiment ($r_{(94)} = 0.85, p < 0.001$; Supplementary Figure S4a). Also in line with the online experiment, fitting the acoustic, MIR and Soundgen models to the data of the lab experiment led to similarly low prediction (Supplementary Tables S4 and S5), that is, in both cases less than 3% of the variance in liking

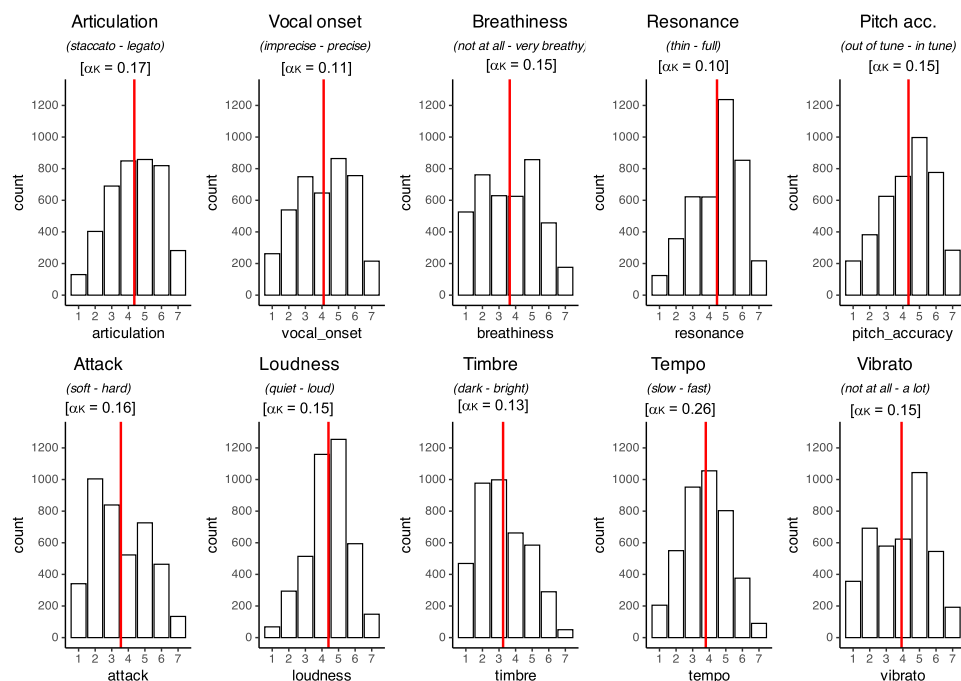


Figure 2. Distribution of perceptual ratings collected in Experiment 2 ($N = 42$ participants), along with anchor words for each bipolar scale. The red line represents the mean rating, and numbers within brackets are Krippendorff's alpha inter-rater agreement based on within-participant normalized ratings (see Supplementary Table S6 for values of Krippendorff's alpha based on raw ratings).

ratings could be explained based on fixed effects, and a larger proportion of variance was captured by random intercepts for participants and stimulus items.

The perceptual model (Fig. 3 and Supplementary Table S7) accounted for 43.4% of the variance in liking ratings based on fixed effects (marginal $R^2=0.433$, conditional $R^2=0.588$). This model includes as fixed effects the 10 collected perceptual features, melody, interactions between melody and articulation, pitch accuracy, resonance and vibrato; and random intercepts for singer, stimulus items (nested in singer) and participants, as well as random slopes for the effect of melody over participants. The coefficients of the perceptual model show that liking ratings were higher for stimuli perceived as more accurate in terms of pitch, with a full resonance, with some vibrato, with precise and soft vocal onsets and with somewhat fast tempi. There were interactions between melody and some of these features: the effect of pitch accuracy (preference for stimuli perceived as in tune) was more pronounced for the melody *Over the rainbow* (possibly because of the sustained octave jump in the beginning of the melody, making pitch inaccuracies particularly easy to detect). The effect of resonance (preference for “full” resonance) was more pronounced for the melody *Don't worry be happy*. The effect of vibrato (preference for use of vibrato) was less pronounced and actually not significant for *Over the rainbow*, indicating a preference for use of (some) vibrato only for the *Don't worry be happy* melody. Regarding the articulation effect (staccato–legato), its interaction with melody indicates preferences in opposite directions depending on the melody: liking ratings were higher as articulation ratings increased (that is, as stimuli were perceived as more legato) for *Over the Rainbow*, and in the opposite direction (that is, as stimuli were perceived as more staccato) for *Don't worry be happy*. This is an interesting example of participants' higher-level stylistic conceptions of how each melody should be performed influencing their liking ratings.

Definition of perceptual features

The discrepancy between the role of perceptual features and the quasi-non-existent role of acoustic features in liking of singing was surprising. To further understand how listeners rate perceptual features, we examined: (1) the inter-rater agreement for each scale; (2) the relationship between the perceptual and acoustic features, independently of the liking ratings; and (3) the role of participants' characteristics in the use of acoustic and perceptual features.

Inter-rater agreement

Inter-rater agreement in participants' perceptual ratings was low, as shown by Krippendorff's α (α_K —see Fig. 2, numbers between brackets, and see Supplementary Table S6 for comparison of α_K values based on raw and normalized ratings). This is in line with literature reporting variably low agreement in perceptual rating tasks using music stimuli both with lay listeners^{22,24} and experts²³. See Supplementary Table S6 for intra-class correlation coefficients (ICCs; single random raters, absolute values) as an alternative measure of agreement, to allow for direct comparison with other relevant studies of perceptual ratings of spoken (e.g.,²⁴), and singing (e.g.,²³) voices that also report ICCs. Note that ICC values were similar to α_K reported in the main text. We also explored an alternative correlational measure of agreement, “mean-minus-one” (MM1), a leave-one-out type of agreement measure¹¹ that measures how much individuals agree with the group (see Supplementary Figure S5). This analysis showed that some participants agree a lot more with the average perceptual evaluation than others.

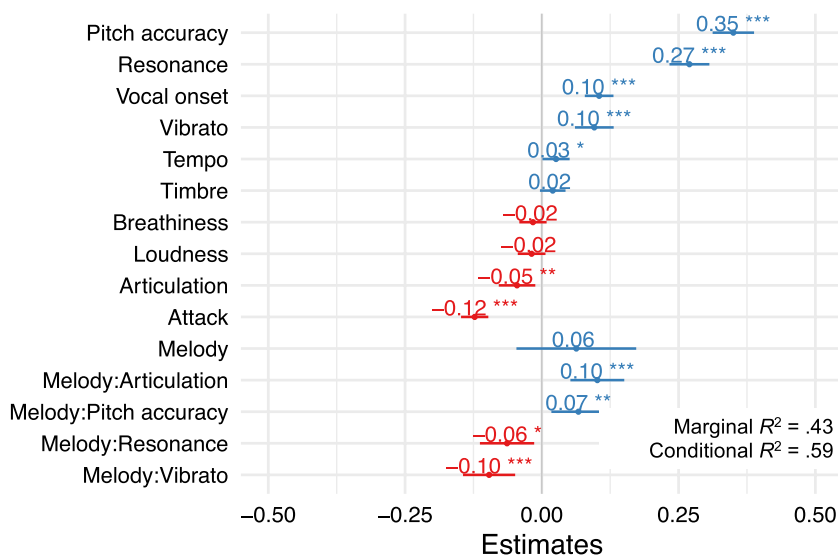


Figure 3. Fixed effects estimates of perceptual model fit on data from Experiment 2 ($N=42$ participants). All predictors are mean-centered and scaled by one standard deviation. Coefficients and their 95% confidence intervals are plotted. Note the group of reference for melody is *Don't worry be happy*. *** $p < 0.001$; ** $p < 0.01$, * $p < 0.05$.

Association between perceptual ratings and acoustic features

Some of the perceptual features are presumably more directly linked to objective properties of the stimuli than others, with timbre (bright–dark) and resonance (thin–full) arguably as the most ‘metaphorical’ judgments participants were asked to make. At the other extreme are pitch accuracy, tempo, and loudness, which are directly related to acoustic measures of pitch interval deviation (relating to f_0), speed of performance, and sound intensity, respectively. Indeed, even though inter-rater agreement of perceptual ratings was low, we found correlations between the mean perceptual ratings (averaged within each scale across all participants) and their corresponding acoustic parameters ranging from $|.21|$ to $|.68|$ (Supplementary Table S8). Note that the perceived amount of vibrato correlated with measurements of vibrato extent (that is, measurement of how wide the vibrato was), and not with measurements of vibrato rate (that is, how fast the vibrato was). We also explored individual differences in this association, by computing the same correlations but based on subsets of perceptual ratings by each individual participant (see Supplementary Table S9 for summary statistics of the individual correlation scores, and Supplementary Figure S6). The range of these individual correlations was very wide (e.g., for tempo, they ranged from -0.20 to 0.63). Importantly, in all cases where we describe an association between mean perceptual ratings and acoustic measurements, there were certain individuals for which correlations with acoustic measurements were higher (or lower) than the overall correlations with mean perceptual ratings. For pitch interval deviation, for instance, the overall correlation with mean perceptual ratings was -0.20 , but for one participant it reached -0.39 . We refer to these individual correlation scores between a participant’s perceptual ratings and the corresponding acoustic measurement as “acoustic sensitivities”, and used these values in exploratory analyses in relation to other participants’ characteristics. Interestingly, we found that participants’ “acoustic sensitivity” correlates with their individual MM1 values in the corresponding scale (i.e., with how much the individual agrees with the group in that particular scale): for loudness: $r(39) = 0.50$; tempo: $r(39) = 0.56$; vibrato: $r(39) = 0.75$; breathiness: $r = -0.82$ (all $ps < 0.001$); but for pitch accuracy this did not reach significance: $r(39) = -0.18$, $p = .272$). Even though the relationship between acoustic measurements and perception is reportedly not that straightforward (see, for instance,^{23,27,28}), it makes sense that participants agreed more with the group perceptual evaluation when their perceptual ratings were more correlated with the corresponding acoustic measurements (or when they were more “correct”). Accordingly, we also found a positive association between participants’ music sophistication, as measured by the Gold-MSI²⁹, and participants’ “acoustic sensitivity” for vibrato and breathiness. That is, more musically sophisticated individuals also seem to be more “sensitive” or “objective” when rating the amount of vibrato and the degree of breathiness of singing performances. Considering that lay listeners are sensitive to pitch interval deviation²⁰, we expected all participants to rely on pitch interval deviation and thus a limited relationship between pitch accuracy and music sophistication, which appears to be the case, with a correlation not reaching significance—though it is in the expected direction (-0.19).

Individual differences in singing appreciation

To explore individual differences in the role of acoustic and perceptual attributes of the stimuli in singing voice appreciation, we fit individual models for each participant from the lab experiment using multiple linear regression (lm function in R). As before, one model (the acoustic model) includes the acoustic features and one (the perceptual model) includes the perceptual features as predictors. As can be seen in Fig. 4, the resulting adjusted R^2 values obtained for each participant (referred to here as individual level of prediction) were larger for the perceptual model than the acoustic model (acoustic model: range from -0.07 to 0.44 , $M = 0.09$, $SD = 0.10$; perceptual model: range from 0.16 to 0.85 , $M = 0.6$, $SD = 0.17$; paired t -test $t(41) = -19.9$, $p < 0.001$).

We also conducted exploratory analyses examining the relationship between the individual level of prediction achieved with the perceptual model (as indicated by the individual adjusted R^2 values resulting from the previous analysis) and participants’ characteristics. To do so, we fit a multiple regression model predicting the individual level of prediction achieved with the perceptual model—that is, the (Fisher z -transformed) adjusted R^2 values calculated for each participant—from the following participant information, collected via questionnaires: general music sophistication, as measured by the Gold-MSI²⁹; the Big-5 personality traits, as measured by the ten-item

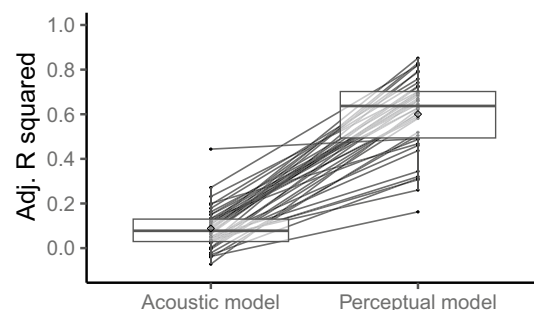


Figure 4. Distribution of adjusted R^2 values of the acoustic and perceptual models fit individually for each participant of the Experiment 2 ($N = 42$ participants). Lines connect values obtained for each participant; diamonds depict the mean value of each distribution. Lower and upper hinges correspond to the first and third quartiles, and whiskers extend from the hinge to $1.5 \times$ inter-quartile range (Tukey-style boxplot).

personality inventory (TIPI;³⁰ in a German version³¹); music preferences in terms of the five dimensions of the MUSIC model, as measured by the STOMP-R³²; participants' average individual MM1 value (across the 10 perceptual scales); participants' average "acoustic sensitivity" (across the five selected scales); and participants' age. The final, reduced model (Supplementary Table S10; see "Methods" section and accompanying .Rmd files for details) indicated that the level of prediction obtained with the perceptual model was higher for more musically sophisticated participants, as well as for older participants; and it was lower for participants with a preference for Mellow music (that is, these participants based their liking ratings less on perceptual attributes of the voices). See Supplementary Table S11 for summary statistics and Supplementary Figure S7 for a correlation matrix of participants' characteristics.

Discussion

We investigated singing voice preferences by attempting to separate the relative contribution of objective stimulus features (estimated through acoustic analyses) from (subjective) internal processes on the listener's side. We asked participants how much they liked singing performances, and modelled liking ratings based on acoustic and perceptual features of the stimuli. We found large individual differences in participants' preferences, as shown by the highly spread distribution of liking ratings for each stimulus and by the low inter-rater agreement. Nevertheless, even though both experiments differed in terms of setting (online or in the lab) and sampled population (USA-based or Germany-based), we found that mean liking ratings between experiments were highly correlated, suggesting that, at least to some extent, the average preferences that emerged were in fact grounded in attributes of the stimuli.

The modeling analysis showed that the predictors included in the acoustic model, although carefully selected based on previous related research, could not explain the variance in liking ratings. Note that (spoken) voice attractiveness has been strongly linked to acoustic features of the voices—for instance, Babel et al.¹⁶ could predict 54% of the variance in voice attractiveness ratings based on acoustic parameters (f_0 , formants, jitter, shitter, CPP, energy and tilt measures) of female voices. In our study, even adding a large number of low-level features from Music Information Retrieval and from the Soundgen package (dedicated to human nonverbal vocalizations) as predictors did not increase the proportion of variance explained. Of course, we did not exhaustively explore all relevant features of the stimuli that might play a role in participants' aesthetic appreciation, but MIR and Soundgen are assumed to provide a wide range of features, which have been successfully used in different settings (e.g.:^{33–39}).

On the other hand, the perceptual model, based on ratings of different perceptual attributes of the voice stimuli, accounted for around 43% of the variance in liking ratings. In other words, the acoustic features themselves did not play a significant role in liking ratings, while the perception that listeners had of acoustic characteristics did. Admittedly, some of the perceptual scales we presented to participants were metaphorical and presumably demanded more top-down processing than others; hence, they could (arguably) be more affected by individual differences in participants' internal processing (e.g.: expertise) than others. Extreme examples of this would be the scales for resonance (thin–full) and timbre (dark–bright), which demanded a higher level of abstraction than scales directly related to a particular physical dimension, like pitch accuracy (out of tune–in tune), tempo (slow–fast) and loudness (quiet–loud), which relate perceptually to f_0 accuracy, speed of performance and sound intensity, respectively.

The associations between acoustic measurements and corresponding average perceptual ratings were between low ($r = -.21$ for pitch accuracy) and moderate ($r = -0.61$ for breathiness). This is in line with previous research showing variable alignment between acoustic/MIR descriptors and human perceptual ratings^{22,23,40,41}. Inspection of individual correlation scores (referred to here as "acoustic sensitivities") shows a wide range of variation across participants, with some of them showing higher (or lower) individual scores than what is reported for the group as a whole. Exploratory analysis showed that individuals with higher "acoustic sensitivity" also agreed more with the average group evaluation (as measured by individual MM1 scores); and that individuals with higher "acoustic sensitivity" for perception of breathiness and vibrato in singing performances scored higher on music sophistication, as measured by the Gold-MSI. These findings highlight the role of individual differences in perception. On the other hand, the limited alignment between perceptual ratings and acoustic estimates may also be related to the limited variance of the acoustic properties themselves in the stimulus set. For instance, the singing performances (by highly trained singers) had remarkably good pitch accuracy: the average pitch interval deviation across all performances was 22 cents ($SD = 11$, $min = 5.4$, $max = 55$), which is under the threshold of mistuning perception for a large part of the population²⁰. In the case of loudness, to avoid possible confounds of loud performances overly influencing liking of our singing stimuli^{42,43}, we normalized all singing performances to the same level of -14 Loudness Units relative to Full Scale (LUFS). So, similarly as for pitch interval deviation, the narrow range of loudness variation in the stimuli likely limited participants' performance in detecting these not so drastic differences across stimuli.

We found low inter-rater agreement both in liking ratings of both experiments and in perceptual ratings of Experiment 2. The low inter-rater agreement could, to some extent, be related to the previously mentioned relative homogeneity of the stimulus set: it would be easier for participants to agree in their pitch accuracy ratings if there were big differences in pitch accuracy between performances. The low inter-rater agreement could also potentially be linked to participants' confusion about what the terms meant, though participants were thoroughly instructed about the definition of each perceptual feature, and studies suggest that untrained listeners are able to evaluate spoken⁴⁴ as well as singing voices²⁴ if suitable scales are made available to them. For instance, using the Geneva Voice Perception Scale (GVPS⁴⁴), listeners without any specific training were able to rate (spoken) voices in terms of loudness, pitch, sharpness, intonation, roughness, articulation, among others, though with sizeable differences in reliability estimates across scales (single measure ICCs ranging from 0.31 for

roughness to 0.85 for loudness). Merrill²⁴ investigated untrained listeners' abilities to rate singing voices from selected excerpts of popular songs, and reported an average ICC measure of 0.39 for nine relevant features, again with sizeable differences across scales ranging from negative values for pitch (−0.44) and articulation precision to higher values for noise (0.54) and tension (0.57). Interestingly, the scales introduced by these studies with untrained listeners led to higher inter-rater agreement than what was reported by Lange and Frieler²³ among experts (sound engineers) rating perceptual features of contrasting music stimuli, with a mean (in the case, also comparable to ICCs) α_K of 0.28 for 16 perceptual variables and a mean α_K of 0.19 for six emotion-related perceptual variables. Generally, individual differences in the use of (subregions of) scales may also drive some of the lack of consistency across raters (i.e., potentially different profiles of raters, see²²). We found that normalizing ratings within participant reduced the variance captured by random intercepts in the linear mixed models considerably, though with no impact on the achieved prediction based on fixed effects.

Music preferences are affected by listeners' cultural backgrounds and social influences such as the projection of one's identity^{45,46}, and are influenced by age⁴⁷, expertise⁴⁸ and past experience⁴⁹. Further, music preferences depend on and convey personality traits^{32,50}, and recent studies have linked psychological traits and preferences for certain low- and high-level musical features^{51–53}. In light of this broad literature, we also explored the influence of participants' characteristics on their use of perceptual features to evaluate singing performances, and found that prediction of liking based on perceptual features was higher among more musically sophisticated and older participants, and lower among participants with a preference for Mellow music (which, according to the MUSIC model³², is slow, quiet and not distorted; associated with the music genres soft rock, R&B and adult contemporary; and perceived as romantic, relaxing and not aggressive). One could speculate that the higher prediction of liking based on perceptual features among more musically sophisticated participants might be linked to higher perceptual skills and/or higher reliance on such skills.

One limitation of these brief explorations comes from the relative (musical) homogeneity of our stimulus set, which limits the opportunity for large perceptual differences in the music material (i.e., only two contrasting melodies). Note that this relative homogeneity from a broader musical point of view was a conscious choice on our part, aimed at emphasizing the variability in the voices in the set. We used high quality, enjoyable naturalistic singing performances, with 16 singers providing three takes for each of two contrasting melodies, encouraging participants to focus primarily on perception of the voices themselves rather than the melodies. Additionally, we made the decision not to (attempt to) record amateur or untrained singers due to the challenges associated with satisfactorily performing these melodies a cappella. Including less polished performances could increase the variability in the stimulus set and the generalizability of findings (see⁵⁴ for an example of a large-scale online approach to collect vocal productions cross-culturally and in the general population). In any case, our finding of consistent average preferences for certain singers across two experiments involving participants from two different countries indicates that there was sufficient variability in the stimulus set to elicit considerable preferences. One further limitation of our stimulus set is that both melodies were likely well known to participants. As an example, the interaction between melody and articulation ratings (indicating an overall preference for staccato performances for the melody Don't worry be happy, and for legato performances for the melody Over the rainbow) suggests that participants had certain expectations on how these melodies should be performed. Familiarity effects⁴⁹ may have influenced participants' enjoyment of performances in different and unexpected ways. In future studies, it would be desirable to investigate aesthetic preferences with a more varied stimulus set and using novel melodies to control for such familiarity effects, and to increase the generalizability of findings.

While the aesthetic appreciation of singing performances can certainly be approached from the perspective of their musical features, it is also subject to specificities of the voice as a special type of stimulus: beyond supporting speech in interpersonal communication, a person's voice can also indicate personality traits^{55–57} and emotional states⁵⁸. The conceptual structure of the aesthetic appreciation of voices differs from conceptual representations underlying other domains of aesthetics⁵⁹: the category of beauty is also prevalent, but not among the most frequent adjectives used; and the three dimensions determinant of voice preferences described in that study—pitch (deep/high), loudness (quiet/loud) and vocal quality (soft/rough)—were not found in studies on the conceptual structure of aesthetics of objects, music or visual arts. Considering the wider role of the voice as a sociobiological signal, it would be interesting to further explore the similarities and specificities of the aesthetic appreciation of spoken and singing voice in an integrated way. For instance, Valentova et al.⁶⁰ reported high correlations between attractiveness ratings of matched spoken and sung voices by naïve participants singing excerpts of *Happy birthday* and national anthems, and even argued that singing and speaking may serve as redundant “backup signals”, both cuing body traits and reproductive fitness. Further studies could expand these finding by investigating the aesthetic appeal of a broad range of contrasting but matched vocalizations (i.e., produced by the same voices).

The famous saying “beauty is in the eye of the beholder”—or perhaps “beauty is in the auditory system of the listener”—illustrates a long-standing debate about the relative contribution to aesthetic appraisals of stimulus features versus processing on the listener's side. Our finding that mean liking ratings correlated highly between participant samples from Germany and the USA hints at cross-cultural average preferences, presumably based (to some extent) on attributes of the stimuli. However, one should keep in mind that both these samples come from Western, educated, industrialized, rich, and democratic (WEIRD⁶¹) nations. It would be desirable to investigate singing voice preferences in other populations, ideally including non-WEIRD populations (as well as using stimuli material in other styles and/or from multiple cultures to increase the generalizability of findings). So far, our results suggest that, in the case of the singing voice, preferences depend substantially on both attributes of the performances and on listeners' characteristics, but this relationship is a complex one: liking ratings do not rely on acoustic features of the voices per se but on how the voices are perceived by participants.

Materials and methods

Material used in both experiments

Recordings of pop singing

Singers. Sixteen highly trained female singers whose age ranged from 19 to 56 years old, with an average of 26.18 ($SD=9.63$), were invited at the Berklee College of Music in Boston, USA. The average length of vocal training among participants was 9.47 years (range: 1 to 29 years, $SD=7.12$), while the average length of musical instruments training was 11.77 years (range: 0 to 30 years, $SD=7.8$).

Selected songs. Each singer was asked to perform the songs *Don't Worry Be Happy*, by Bobby McFerrin, and *Over the Rainbow*, by Harold Arlen, without lyrics and in a /u/ sound. Each singer recorded between four and seven takes of each melody. For each melody, we selected three takes from each of the 16 singers for acoustic analysis and subsequent use in the experiment. The stimulus material thus comprised 48 excerpts of each melody, for a total of 96 stimuli, all lasting between 6 and 9 s.

Recording procedure. After receiving instructions and signing the consent form, singers were invited to warm up their voices for 3 to 5 min. They were then asked to sing the theme of *Don't Worry Be Happy* (which is whistled in the original song—Fig. 1A) five times. They were then asked to sing the first phrase of the refrain of *Over the Rainbow* five times (Fig. 1A). The starting pitch height (E5 or E4) was played before each recording. Recordings took place in a sound attenuated booth, using a Neumann (U87) microphone placed at approximately one meter in front of the singer's mouth; with 16 bits per sample and 48 kHz sampling rate.

Audio processing and stimulus selection. For each singer, we selected three takes of each melody for further processing and use in the experiment. Stimuli were cut using Audacity software (version 2.3.3). All stimuli were normalized to the same loudness level of -14LUFS according to the EBU-R128 standards with the software To Audio Converter (Version 1.0.14(902)).

Acoustic analyses. The procedures to segment performances into individual notes and calculate vibrato rate, vibrato extent and energy distribution are extensively described in^{19,62,63} and are described succinctly here. Using AudioSculpt 2.9.4v3 software (IRCAM, Paris, France), we first manually placed markers on boundaries of individual notes and used this information to segment each singing performance into individual notes with a sox script. Acoustic analysis was then conducted based on individual notes. Using OpenMusic 6.3 software (IRCAM, Paris, France), we focused on one note per performance (the last note of the *Don't Worry Be Happy* excerpt, a F3 sharp, and the first note of *Over the Rainbow* excerpt, an E3) to calculate vibrato rate, vibrato extent and energy distribution. Vibrato rate corresponds to the number of quasiperiodic modulations of the f_0 per second (in Hz). Vibrato extent corresponds to the amplitude of the f_0 variations within the same tone (in cents). The energy distribution measure is the ratio of the energy of the 2.4–5.4 kHz band divided by the energy across the 0–10 kHz band—the rationale behind this computation is that a high score in the energy distribution variable indicates a strong reinforcement of the band containing the singer's formant. We used Praat⁶⁴ (Version 6.0.46), to extract: mean f_0 ; jitter (local), the perturbation in the f_0 from cycle to cycle; and shimmer (local), the perturbation in the amplitude of the f_0 from cycle to cycle. Additionally, features relevant to the description of spoken voice were extracted using VoiceSauce⁶⁵, following Babel et al.¹⁶ and Bruckert et al.¹⁷: spectral tilt, a measure of voice quality, where higher values of tilt indicate breathier voices and lower values of tilt indicate creakiness⁶⁶. We computed the amplitude of the first harmonic minus the amplitude of the second harmonic (H1–H2), and the longer distance measure of the first harmonic minus the peak amplitude of the first, second, and third formants (H1–A1, H1–A2 and H1–A3, respectively, both corrected and uncorrected, exploratorily). We also extracted the Harmonic-to-Noise ratio (the ratio between periodic and non-periodic components of a speech sound; the higher the HNR, the less hoarse a voice sounds⁶⁷ in the 0–3.5 kHz range (HNR35); and the Cepstral Peak Prominence (CPP), a measure of breathiness and overall dysphonia⁶⁸. To measure how accurate performances were in terms of pitch, based on mean f_0 of each individual note, we converted f_0 values from Herz to cents (100 cents corresponds to one semitone; the reference lowest note used was 261.626 Hz); and then computed successive pitch interval deviations by subtracting the f_0 of adjacent notes and comparing values to “correct” pitch intervals (i.e., according to musical notation). For example, if the interval between two successive notes of a melody should form a major second (which corresponds to an interval of 200 cents), and a singer made a quarter-tone mistake, i.e. if she sang an interval of 150 or 250 cents, we considered that the error was 50 cents in relation to the expected interval. We finally averaged the pitch interval deviation per performance. This procedure has been used and validated in several previous studies^{19,21,62,69}. Note that we noticed imprecision in estimates of mean f_0 for some very short notes, so we excluded the fourth note of *Over the rainbow* and the seventh note of *Don't worry be happy* from pitch interval deviation computations. Please see Supplementary Figure S8 for a demonstration of this. For jitter and shimmer, we noticed numerous aberrant values for very short notes, indicating measurement imprecision on such short notes, so we trimmed values higher than two standard deviations above the mean value (for each of these two features). This meant excluding 5% of all measurements for shimmer and 2.86% for jitter. We also computed average tempo of each performance as the excerpt's length divided by the number of beats. A summary of the acoustic properties of the voice stimuli used in the study is presented in Supplementary Table S1 (also see Supplementary Figure S2 for boxplots of these measures). A correlation matrix of acoustic features included in the acoustic model and liking ratings from both experiments is reported in Supplementary Figure S9.

Batch extraction of further acoustic/MIR features. To further characterize our stimulus set and based on whole performances, we also analyzed our stimuli using features from Soundgen²⁵, an open-source toolbox for voice synthesis, manipulation, and analysis, published as a library (package) for the R programming language; MIRToolbox⁷⁰, a Matlab toolbox for music information retrieval (MIR); and Essentia⁷¹, an open-source C++ library for music information retrieval. From Soundgen, we used the `analyze` function to batch extract several acoustic features. From the output of 156 features (presented as mean, median, and standard deviation summaries per file), we excluded median statistics and `voiced_` alternatives because they were highly correlated with their equivalent mean statistics and general correspondent measures, respectively. This led to a core subset of 54 features: `amEnvDep_mean`, `amEnvDep_sd`, `amEnvFreq_mean`, `amEnvFreq_sd`, `amMsFreq_mean`, `amMsFreq_sd`, `amMsPurity_mean`, `amMsPurity_sd`, `ampl_mean`, `ampl_sd`, `CPP_mean`, `CPP_sd`, `dom_mean`, `dom_sd`, `entropy_mean`, `entropy_sd`, `entropySh_mean`, `entropySh_sd`, `flux_mean`, `flux_sd`, `fmDep_mean`, `fmDep_sd`, `fmFreq_mean`, `fmFreq_sd`, `harmEnergy_mean`, `harmEnergy_sd`, `harmHeight_mean`, `harmHeight_sd`, `HNR_mean`, `HNR_sd`, `loudness_mean`, `loudness_sd`, `novelty_mean`, `novelty_sd`, `peakFreq_mean`, `peakFreq_sd`, `pitch_mean`, `pitch_sd`, `quartile25_mean`, `quartile25_sd`, `quartile50_mean`, `quartile50_sd`, `quartile75_mean`, `quartile75_sd`, `roughness_mean`, `roughness_sd`, `specCentroid_mean`, `specCentroid_sd`, `specSlope_mean`, `specSlope_sd`, `subDep_mean`, `subDep_sd`, `subRatio_mean`, `subRatio_sd`. For the two other libraries of MIR features (MIRToolbox and Essentia), we focused on low-level features, since higher-level features would not vary significantly across our two types of short a cappella melody excerpts. From the MIRToolbox, we extracted the following features: `low_energy`, `spec_entropy`, `brightness`, `flatness`, `zerocross`, `pitch`, `roughness`, `mirtempo`, `rolloff85`, `rms`, `rolloff95`, `spread`, `skewness`, `regularity`, `flux_med`, `keyclarity`, `flux`, `kurtosis`, `mode`, `subband (1–10)`, `mfcc (1–10)`, `centroid`, `pulse_clarity`, `harmonic_change`, `fluctuation_max`, `spectral_novelty`. Note that `subband` and `mfcc` refer to 10 spectral subbands. For each feature, an average value and standard deviation were extracted, adding up to a total of 80 variables (since for some std variables extraction failed, as well as for `mfcc1_mean` and `mfcc_std`). From Essentia, we used the out-of-box executable streaming extractor `freesound` (https://essentia.upf.edu/freesound_extractor.html) to extract the following low-level features: `average_loudness`, `barkbands_kurtosis`, `barkbands_skewness`, `barkbands_spread`, `dissonance`, `hfc`, `pitch`, `pitch_instantaneous_confidence`, `pitch_salience`, `silence_rate_20dB`, `silence_rate_30dB`, `silence_rate_60dB`, `spectral_complexity`, `spectral_crest`, `spectral_decrease`, `spectral_energy`, `spectral_energyband_high`, `spectral_energyband_low`, `spectral_energyband_middle_high`, `spectral_energyband_middle_low`, `spectral_entropy`, `spectral_flatness_db`, `spectral_flux`, `spectral_rms`, `spectral_rolloff`, `spectral_skewness`, `spectral_spread`, `spectral_centroid`, `spectral_kurtosis`, `spectral_strongpeak`, `zerocrossingrate`, `barkbands (01–27)`, `frequency_bands (01–27)`, `gfcc (01–13)`, `mfcc (01–13)`, `scvalleys (01–06)`, `spectral_contrast (01–06)`. Note that `barkbands` and `frequency_bands` refer to 27 spectral subbands, `gfcc` and `mfcc` refer to 13 spectral subbands, and `scvalleys` and `spectral_contrast` to six subbands. For each feature, an average value and standard deviation were extracted, and in some cases also the variance of the derivative of a feature (`dvar`), adding up to a total of 327 features. Note that features were highly correlated with each other: see Supplementary Information (Supplementary Methods: Selecting features from Music Information Retrieval) and accompanying `.Rmd` files for details and strategies used to subset features and subsequently model liking ratings based on Soundgen and MIR features.

Methods specific to experiment 1

Participants

Recruited from Amazon Mechanical Turk (<https://www.mturk.com>), a total of 330 participants (180 male, 147 female, 3 undisclosed, $M = 38.3$ years old, $SD = 11.6$) completed the whole experiment online. All participants provided informed consent in accordance with the Max Planck Society Ethics Council approved protocol (application 2018–38) and were paid at a US \$9/hour rate according to how much of the experiment they completed.

Procedure

Experiment 1 was designed and implemented using PsyNet (<https://www.psynet.dev>), a Python package for performing complex online behavioral experiments at large scale⁷². PsyNet is based on the Dallinger framework (<https://dallinger.readthedocs.io>) for hosting and deploying experiments. Participants interact with the experiment via a web browser, which communicates with a back-end Python server cluster responsible for organizing the experiment and communicating with participants. The server cluster was provisioned using Heroku.

Audio pre-screening task. Before starting the rating task, participants completed an audio pre-screening task⁷³ to ensure they were wearing headphones and could perceive subtle sound differences. This three-alternative forced-choice task consisted in identifying the quietest of three tones. These tones are constructed to elicit a phase cancellation effect, such that when played on loudspeakers, the order of quietness changes, causing the participant to respond incorrectly. Participants passed the test only after answering correctly to at least four of the six trials.

Liking rating task. Participants were asked to rate how much they liked singing performances. In each trial, a melody excerpt was played, followed by a screen with the question: How much did you like this performance? Participants responded with a mouse on a graphic display of a scale ranging from 1 (not at all) to 9 (a lot). The next trial began immediately after this response. Each participant rated two blocks of performances, one for each melody, in counterbalanced order. Each block was composed of 16 performances of one melody, each performed by a different singer, followed by 8 repetitions, that were used exclusively in a test–retest consistency analysis.

Participants' information. After the two blocks of trials, participants answered 15 questions from a shortened scale of General Music Sophistication of the Goldsmiths Music Sophistication Index, (items: AE_01, AE_02, EM_04, MT_02, MT_03, MT_06, MT_07, PA_04, PA_08, SA_01, SA_02, SA_03, SA_04, SA_05, SA_06)²⁹. Finally, participants were asked whether they liked the experiment (open field answers). Answers were mostly positive, with 95.8% of the participants answering “Yes”, 0.9% answering “No”, 0.9% answering some type of mixed response one could summarize as “partially”, and 2.4% providing no answer.

Statistical analyses

All analyses were performed using R Statistical Software (version 4.1.2)⁷⁴ and R Studio (version 2023.06.0 + 421)⁷⁵. Four participants were not included in further analysis (three participants rated all stimuli “1” and one participant rated all stimuli “8” or “9”), leaving a total of 326 participants.

Intra-rater agreement

To analyze the consistency of participants' ratings, we used the 16 stimuli that were presented twice for each participant (8 from each melody), and calculated a Pearson's correlation score between first and second presentation for each participant.

Normalization of ratings

To account for different rating strategies (i.e., different sections of the scale being used by individual participants), we normalized ratings within each participant between 0 and 1. As expected, the original 9-point scale raw ratings and the normalized ratings were highly correlated ($r = 0.87$, $p < 0.001$). Also note that fitting models based on normalized instead of raw ratings only affected the amount of variance explained by random effects (specifically, captured by subjects' random intercepts), but did not have any meaningful impact on the amount of prediction by the fixed effects: fitting the same acoustic and MIR models to predict raw liking ratings instead of (within-participant) normalized liking ratings produced similar marginal R^2 values (please see Supplementary Table S12 for this comparison).

Inter-rater agreement

In order to examine inter-rater agreement, we used Krippendorff's alpha (α_K)^{76,77}, a generalization of several known reliability indices, where an $\alpha_K = 1$ indicates perfect agreement and $\alpha_K = 0$ indicates no agreement at all. We used the `kripp.alpha` function from the `irr` R package⁷⁸.

Predicting liking from acoustic descriptors

Different linear mixed effects analyses were proposed using the `lmer` function from the `lme4` package⁷⁹. For all models reported, residual plots and QQ-curves were visually inspected to make sure there were no deviations from normality or homoscedasticity, and p -values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. Variance inflation factor (VIF) was under 5 for all predictors kept in the models. Predictors were mean-centered and scaled by one standard deviation before model fitting. Note that the analyses of the online experiment included the 32 trials per subject (16 from each melody) but not the repeated trials, which only served the purpose of consistency analysis. The null model predicted (within-participant normalized) liking ratings and included only random effects (random intercepts for subjects and stimuli items nested in singers). The acoustic model also included as fixed effects predictors shown to be relevant in studies of perception of pitch accuracy^{19,21,80} and voice attractiveness¹⁷: pitch interval deviation, vibrato extent, vibrato rate, energy ratio, tempo and harmonics-to-noise ratio, jitter, shimmer and the voice quality measure CPP. Note that the tilt measures H1–H2, H1–A1, H1–A2 and H1–A3 were not included in the model due to multicollinearity issues. A second set of linear mixed models, referred to as MIR models, was fit to predict (within-participant normalized) liking ratings from subsets of low-level MIR features, with subjects and stimuli items (nested in singers) as random effects. Since MIR features were highly correlated with each other, we first explored three different approaches to subset them and then fit three linear mixed models based on the resulting subsets of features. We also fit one separate model based on features from the `Soundgen` toolkit. Please see the Supplementary Information (Supplementary Methods: Selecting features from Music Information Retrieval/ Selecting features from the `Soundgen` package) and accompanying `.Rmd` files for details about feature reduction and model selection.

Method specific to experiment 2

Participants

Forty-two participants (26 female, 16 male), aged from 22 to 75 years old ($M = 36.8$, $SD = 16.1$), completed the whole experiment. The experimental procedure was ethically approved by the Ethics Council of the Max Planck Society, and was undertaken with written informed consent of each participant.

Procedure

The experiment was implemented in Labvanced⁸¹. Participants were tested on a computer located in the laboratories of the Max Planck Institute for Empirical Aesthetics in Frankfurt, Germany. The liking rating scale and the perceptual evaluation scales were presented on the same screen, in German, together with the experimental stimuli to rate. As in the online experiment, participants rated how much they liked each singing performance on a scale from 1 (not at all) to 9 (a lot).

Perceptual ratings. Particular to Experiment 2 were perceptual ratings: we developed a series of scales relative to perceptual attributes shown to be relevant in the appreciation of singing. Perceptual attributes were rated on bipolar scales ranging from 1 to 7 and displaying contrasting anchor words on each pole. The scales, their bipolar anchors and the definition presented to participants were the following (translated to English here): Pitch accuracy (in tune–out of tune): how precise is each note along the melody—is the performance in-tune or out-of-tune? Loudness (soft–loud): the magnitude of the auditory sensation: is the voice quiet or loud? Tempo (slow–fast): the speed or pace of the performance—is the performance slow or fast? Articulation (staccato–legato): how notes are connected to each other—are notes detached (staccato) or connected (legato)? Breathiness (not at all–a lot): the amount of air flow in the voice—how breathy does the voice sound? Resonance (thin–full): the fullness or reverberation of a voice: how full is the voice? Timbre (dark–bright): the perceived sound quality of the voice: does the voice sound dark or bright? Attack/Voice onset 1 (soft–hard): the way in which a note begins—is the start of notes soft or hard? Voice onset 2 (precise–imprecise): the way in which a note begins—is the start of notes precise? Vibrato (not at all–a lot): A slight and periodic oscillation of the pitch of a sustained note—how much vibrato does the performer use? Please see Supplementary Table S13 for the German version of these scales and instructions used in the experiment. There was also one question about the vocalic sound produced by the singer, because even though they were asked to sing in an /u/ sound, in many cases the produced vowel wasn't that clear or constant, and we hypothesized this could influence participants' liking. Therefore, participants also had to indicate what was the dominant vowel sound perceived in each stimulus in a forced choice response. Participants indicated perceiving a /u/ sound in 53.4% of trials; a /a/ sound in 16.2% of trials; a /o/ sound in 10.8% of trials; a /i/ sound in 1.54% of trials; and an unclear vowel sound in 18% of trials. Note that this predictor was later found insignificant during the modeling analysis and dropped from the perceptual model.

The experiment was divided into six blocks (three per melody). Half of the participants started with the three blocks for *Over the Rainbow*, half with the opposite order. Each block comprised 16 trials, corresponding to one take by each of the 16 singers, presented in a randomized order. The order of these three blocks within each melody was counterbalanced across participants. Following the general instructions, we presented the definitions of the scales and five examples they could listen to. For each experimental stimulus, participants could click on the “play” button as many times as they wanted to and listen to the stimulus again as they rated 10 bipolar scales, the scale for liking and the forced-choice question for dominant vowel. The next page was proposed and viable when all scales and the vowel question were completed and participants pressed the “next” button. Half-way through the experiment (i.e., between the 3 blocks of the first and the second melody), participants completed three questionnaires (see below). Participants completed the experiment at their own pace and took between 1.5 and three hours to complete the experiment. They could take a break at any desired moment, and additional breaks were included between blocks of trials.

Participants' information. In addition to providing biographical data, participants completed three questionnaires:

- 1) The 18-items subscale of Music Sophistication from the Goldsmiths Music Sophistication Index²⁹ in German, as computed in the Gold-MSI configurator (<https://shiny.gold-msi.org/gmsiconfigurator>; items included: AE_01, AE_02, AE_05, AE_07, EM_04, MT_01, MT_02, MT_03, MT_06, MT_07, PA_04, PA_08, SA_01, SA_02, SA_03, SA_04, SA_05, SA_06)
- 2) The Ten-Item Personality Inventory (TIPI), a short self-report measure of the Big-Five personality domains³⁰, in the German version³¹. Each of the five personality dimensions—Extraversion, Agreeableness, Conscientiousness, Emotional stability (or Neuroticism) and Openness to new experiences—was measured by two items, selected from the high and low poles of each domain. Each question presented two central descriptors, and participants had to rate on a scale from 1 (disagree strongly) to 7 (agree strongly) how much those two traits applied to them.
- 3) The Reviewed Short Test of Music Preference (STOMP-R), a short self-report inventory for musical preferences³²; see⁸² for a German validation). Rentfrow et al.³² suggested a latent 5-factor structure underlying music preferences, which they named the MUSIC model (Mellow, Unpretentious, Sophisticated, Intense and Contemporary). Participants were asked to indicate their preferences for 23 different music genres on a 7 point scale (dislike strongly, dislike moderately, dislike a little, neither like nor dislike, like a little, like moderately, like strongly). To compute each participant's preference for each dimension (see instructions retrieved from the author's website: <https://www.psd.psychol.cam.ac.uk/projects-measures>), we averaged scores within the following genres: Dance/Electronica, New Age, World (Mellow); Pop, Country, Religious (Unpretentious). Blues, Jazz, Bluegrass, Folk, Classical, Gospel, Opera (Sophisticated); Rock, Punk, Alternative, Heavy Metal (Intense); Funk, Rap/hip-hop, Reggae, Soul/R&B (Contemporary). As in⁴⁷, we excluded the genres Soundtrack and Oldies from analysis because they load on different dimensions. We instructed participants to leave blank the genres they did not know.

Statistical analyses

Normalization of ratings. As in Experiment 1, we normalized all collected ratings within-participant (between 0 and 1) to account for individual differences in the use of subregions of the rating scales. Raw and normalized ratings were highly correlated ($r=0.95$ for liking; $r>0.92$ for all perceptual scales; please see Supplementary Figure S3 for the distribution of normalized perceptual ratings). Note that also for data from Experiment 2, fitting our models based on raw versus normalized ratings leads to similar prediction based on fixed effects, and only meaningfully impacts the variance captured by random effects (please see Supplementary Table S12 for this comparison).

Inter-rater agreement. Once again, we examined inter-rater agreement with α_k , but we also computed intraclass correlation scores (ICC2 or single random raters, absolute values), using the ICC function in the psych R package⁸³, to facilitate comparison with other studies reporting this agreement measure. We also computed “mean-minus-one” agreement. To compute MM1, a Pearson correlation is computed between a given participant’s set of ratings and the average ratings of all other participants. This is done for all participants in the sample. The resulting individual correlations are first converted to z scores (to reduce bias in estimates⁸⁴), averaged and converted back into an r score for easier interpretation of the final MM1 measure. We calculated MM1 agreement for liking and for each of the perceptual scales separately (Supplementary Figure S5), and were not primarily interested in the overall MM1 index per perceptual scale per se, but on the distribution of individual MM1 values (i.e., for each participant): both the individual MM1 scores in each scale and one average value per participant (computed across all 10 perceptual scales) were used in exploratory analyses, in relation to other participant-related data.

Relationship between acoustic measurements and perceptual ratings. We computed Pearson correlations between our acoustic measurements (in some cases taken from the toolboxes MIRToolbox, Essentia and/or Soundgen) and the corresponding average perceptual rating (see Supplementary Table S8). Note that for some of the perceptual scales, there is not one clear acoustic correlate. We computed the correlation between our estimates of pitch interval deviation and participants’ ratings of perceived pitch accuracy; estimates of tempo (as measured by beats per minute) and as perceived by participants; estimates of vibrato extent and perceived amount of vibrato; estimates of vibrato rate and perceived amount of vibrato; estimates of loudness_mean (from the Soundgen package) and perceived loudness (note this measure seems more appropriate than Soundgens’ ampl_mean or MIRToolbox’s rms_mean, since it takes into account the sensitivity of human ears to different frequencies); estimates of cepstral peak prominence (CPP; we compared the estimates from VoiceSauce, which were included in our acoustic model, and from Soundgen) and perceived breathiness; and the measure flux_mean (we compared measures from the MIRToolbox, Soundgen and Essentia) and perceived articulation.

Individual differences in the relationship between acoustics and perception. We also explored individual differences in the relationships between acoustic measurements and individual participants’ perceptual ratings. To do so, we computed the correlations mentioned above, but based on subsets of perceptual ratings by each individual participant—we refer to these as “acoustic sensitivity”. This is illustrated in Supplementary Figure S6 (and see summary statistics of the individual correlation scores in Supplementary Table S9). Further, we conducted exploratory analyses based on participants’ “acoustic sensitivity”. For this, we focused on the most straightforward correlations from Supplementary Table S8: we selected the correlations between measurements of pitch interval deviation and perceived pitch accuracy; tempo in beats per minute and as perceived by participants; loudness_mean (from Soundgen) and perceived loudness; Vibrato extent and perceived vibrato; and CPP_mean (from VoiceSauce, which is a predictor in our acoustic model) and perceived breathiness. We first performed a Fisher z-transformation on the individual r-scores (to reduce bias in estimates⁸⁴). We then used these values in two exploratory analyses: a) we calculated the correlation between these values (separately for each scale) and participants’ individual MM1 values (also as z-scores), in an effort to understand if individuals with higher “acoustic sensitivity” tend to agree more with the average group evaluation; b) we calculated an average “acoustic sensitivity” value per participant (by averaging values across the five selected scales) and included it as a predictor in a model investigating the level of prediction achieved with the perceptual model (see Role of participants’ characteristics).

Predicting liking from acoustic and perceptual features. We performed the same linear mixed effects analyses reported for Experiment 1, that is, we fit the same acoustic and MIR/Soundgen models. In addition, we built a perceptual model predicting liking ratings from the collected perceptual ratings. The final perceptual model includes as fixed effects the 10 collected perceptual features and melody; interactions between melody and articulation, pitch accuracy, resonance and vibrato; and random intercepts for singer, stimuli items (nested in singer) and participants, as well as random slopes for the effect of melody on participants.

Role of participants’ characteristics. To investigate individual differences in how much participants based their liking ratings on acoustic and perceptual features of the stimuli, we proposed two (individual) multiple regression models (with the lm function in R) to each participant, using either the predictors from the acoustic or the perceptual model. We refer to the resulting adjusted R^2 values of these individual multiple regression models as “individual level of prediction” (for each participant). We then used the individual level of prediction based on perceptual features to investigate the role of participants’ characteristics on how much they based their liking ratings on perceptual features of the stimuli. To do so, we entered the individual level of prediction for the perceptual model (that is, the adjusted R^2 values resulting from individual fitting of perceptual models) as the dependent variable in a multiple regression model where the predictors were participant-related data collected in the questionnaires of general music sophistication, musical preferences, and psychological traits, as well as participants’ age, participants’ average individual MM1 value (across the 10 perceptual scales) and participants’ average “acoustic sensitivity” (across five selected scales—see Individual differences in the relationship between acoustics and perception). To improve model fit, we performed a Fisher z-transformation of the individual adjusted R^2 values prior to fitting the model. We used stepwise selection to reduce the model (with the step function in R).

Data availability

Supplementary information, the raw data from both experiments and analyses code (as .Rmd files), as well as examples of the singing performances, can be found at <https://osf.io/ts24m/>. The whole set of singing performances can be made available to interested researchers upon request.

Received: 7 November 2023; Accepted: 3 April 2024

Published online: 18 April 2024

References

- Savage, P. E., Brown, S., Sakai, E. & Currie, T. E. Statistical universals reveal the structures and functions of human music. *Proc. Natl. Acad. Sci.* **112**, 8987–8992 (2015).
- Mehr, S. A. *et al.* Universality and diversity in human song. *Science* **366**, eaax0868 (2019).
- Demetriou, A., Jansson, A., Kumar, A. & Bittner, R. M. Vocals in music matter: the relevance of vocals in the minds of listeners. In *Proceedings of the 19th ISMIR Conference* (Paris, France, 2018).
- Fechner, G.T. *Vorschule Der Aesthetik [Elements of Aesthetics]*. (Breitkopf & Härtel, Weimar, Leipzig: Universitätsbibliothek, 1876).
- McManus, I. C., Cook, R. & Hunt, A. Beyond the Golden Section and normative aesthetics: Why do individuals differ so much in their aesthetic preferences for rectangles?. *Psychol. Aesthet. Creat. Arts* **4**, 113–126 (2010).
- Bar, M. & Neta, M. Humans prefer curved visual objects. *Psychol. Sci.* **17**, 645–648 (2006).
- Rhodes, G., Proffitt, F., Grady, J. M. & Sumich, A. Facial symmetry and the perception of beauty. *Psychon. Bull. Rev.* **5**, 659–669 (1998).
- Tinio, P. P. L. & Leder, H. Natural scenes are indeed preferred, but image quality might have the last word. *Psychol. Aesthet. Creat. Arts* **3**, 52–56 (2009).
- Vessel, E. A. Neuroaesthetics. In *Encyclopedia of Behavioral Neuroscience, 2nd edition* 661–670 (Elsevier, 2022). <https://doi.org/10.1016/B978-0-12-809324-5.24104-7>
- Vessel, E. A., Stahl, J., Maurer, N., Denker, A. & Starr, G. G. Personalized visual aesthetics. In (eds. Rogowitz, B. E., Pappas, T. N. & de Ridder, H.) 90140S (San Francisco, California, USA, 2014). <https://doi.org/10.1117/12.2043126>
- Vessel, E. A., Maurer, N., Denker, A. H. & Starr, G. G. Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition* **179**, 121–131 (2018).
- Vessel, E. A. Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *J. Vis.* **10**, 1–14 (2010).
- Brattico, E. The empirical aesthetics of music. In *The Oxford Handbook of Empirical Aesthetics* (eds. Nadal, M. & Vartanian, O.) 573–604 (Oxford University Press, 2021). <https://doi.org/10.1093/oxfordhb/9780198824350.013.26>
- Wassiliwizky, E. & Menninghaus, W. Why and how should cognitive science care about aesthetics?. *Trends Cogn. Sci.* **25**, 437–449 (2021).
- Brattico, E., Bogert, B. & Jacobsen, T. Toward a neural chronometry for the aesthetic experience of music. *Front. Psychol.* **4**, 26032 (2013).
- Babel, M., McGuire, G. & King, J. Towards a more nuanced view of vocal attractiveness. *PLoS ONE* **9**, e88616 (2014).
- Bruckert, L. *et al.* Vocal attractiveness increases by averaging. *Curr. Biol.* **20**, 116–120 (2010).
- Mook, A. T. & Mitchel, A. D. The role of audiovisual integration in the perception of attractiveness. *Evol. Behav. Sci.* **13**, 1–15 (2019).
- Larrouy-Maestri, P., Magis, D. & Morsomme, D. The evaluation of vocal pitch accuracy. *Music Percept.* **32**, 1–10 (2014).
- Larrouy-Maestri, P., Magis, D., Grabenhorst, M. & Morsomme, D. Layman versus professional musician: Who makes the better judge?. *PLOS ONE* **10**, e0135394 (2015).
- Larrouy-Maestri, P., Morsomme, D., Magis, D. & Poeppel, D. Lay listeners can evaluate the pitch accuracy of operatic voices. *Music Percept.* **34**, 489–495 (2017).
- M. Schedl, H. Eghbal-Zadeh, E. Gómez, & Tkalcic, M. An analysis of agreement in classical music perception and its relationship to listener characteristics. In *Proceedings of the 17th ISMIR Conference* 578–583 (2016).
- Lange, E. B. & Frieler, K. Challenges and opportunities of predicting musical emotions with perceptual and automatized features. *Music Percept.* **36**, 217–242 (2018).
- Merrill, J. Auditory perceptual assessment of voices: Examining perceptual ratings as a function of voice experience. *Curr. Psychol.* <https://doi.org/10.1007/s12144-022-02734-7> (2022).
- Anikin, A. Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behav. Res. Methods* **51**, 778–792 (2019).
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z. & Damer, E. Data quality of platforms and panels for online behavioral research. *Behav. Res. Methods* **54**, 1643–1662 (2021).
- Müllensiefen, D., Pfeleiderer, M. & Frieler, K. The perception of accents in pop music melodies. *J. New Music Res.* **38**, 19–44 (2009).
- Larrouy-Maestri, P. *et al.* Ironic twists of sentence meaning can be signaled by forward move of prosodic stress. *J. Exp. Psychol. Gen.* **152**, 2438–2462 (2023).
- Müllensiefen, D., Gingras, B., Musil, J. & Stewart, L. The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE* **9**, e89642 (2014).
- Gosling, S. D., Rentfrow, P. J. & Swann, W. B. A very brief measure of the Big-Five personality domains. *J. Res. Personal.* **37**, 504–528 (2003).
- Muck, P. M., Hell, B. & Gosling, S. D. Construct validation of a short five-factor model instrument. *Eur. J. Psychol. Assess.* **23**, 166–175 (2007).
- Rentfrow, P. J., Goldberg, L. R. & Levitin, D. J. The structure of musical preferences: A five-factor model. *J. Pers. Soc. Psychol.* **100**, 1139–1157 (2011).
- Casey, M. A. Music of the 7Ts: Predicting and decoding multivoxel fMRI responses with acoustic, schematic, and categorical music features. *Front. Psychol.* **8**, 1179 (2017).
- Fricke, K. R., Greenberg, D. M., Rentfrow, P. J. & Herzberg, P. Y. Computer-based music feature analysis mirrors human perception and can be used to measure individual music preference. *J. Res. Personal.* **75**, 94–102 (2018).
- Giraldo, S. *et al.* Automatic assessment of tone quality in violin music performance. *Front. Psychol.* **10**, 334 (2019).
- Papadopoulos, H. & Peeters, G. Large-scale study of chord estimation algorithms based on chroma representation and HMM. In *2007 International Workshop on Content-Based Multimedia Indexing* 53–60 (IEEE, Talence, France, 2007). <https://doi.org/10.1109/CBMI.2007.385392>
- Tzanetakis, G. & Cook, P. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.* **10**, 293–302 (2002).
- Anikin, A. The link between auditory salience and emotion intensity. *Cogn. Emot.* **34**, 1246–1259 (2020).
- Anikin, A., Canessa-Pollard, V., Pisanski, K., Massenet, M. & Reby, D. Beyond speech: Exploring diversity in the human voice. *iScience* **26**, 108204 (2023).
- Daikoku, H. *et al.* Agreement among Human and Automated Estimates of Similarity in a Global Music Sample. <https://doi.org/10.31234/osf.io/76fmq> (2020)

41. Lee, H., Hoeger, F., Schoenwiesner, M., Park, M. & Jacoby, N. Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms. (2021) <https://doi.org/10.5281/zenodo.5625680>
42. Wang, D., Ziano, I., Hagger, M. S. & Chatzisarantis, N. L. D. Loudness perceptions influence feelings of interpersonal closeness and protect against detrimental psychological effects of social exclusion. *Personal. Soc. Psychol. Bull.* **48**, 566–581 (2022).
43. Welch, D. & Fremaux, G. Understanding why people enjoy loud sound. *Semin. Hear.* **38**, 348–358 (2017).
44. Bänziger, T., Patel, S. & Scherer, K. R. The role of perceived voice and speech characteristics in vocal emotion communication. *J. Nonverbal Behav.* **38**, 31–52 (2014).
45. North, A. C. & Davidson, J. W. Musical taste, employment, education, and global region. *Scand. J. Psychol.* **54**, 432–441 (2013).
46. North, A. C., Hargreaves, D. J. & O'Neill, S. A. The importance of music to adolescents. *Br. J. Educ. Psychol.* **70**, 255–272 (2000).
47. Bonneville-Roussy, A., Rentfrow, P. J., Xu, M. K. & Potter, J. Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. *J. Personal. Soc. Psychol.* **105**, 703–717 (2013).
48. North, A. C. & Hargreaves, D. J. Subjective complexity, familiarity, and liking for popular music. *Psychomusicol. J. Res. Music Cogn.* **14**, 77–93 (1995).
49. Peretz, I., Gaudreau, D. & Bonnel, A.-M. Exposure effects on music preference and recognition. *Mem. Cognit.* **26**, 884–902 (1998).
50. Rentfrow, P. J. & Gosling, S. D. The do re mi's of everyday life: The structure and personality correlates of music preferences. *J. Pers. Soc. Psychol.* **84**, 1236–1256 (2003).
51. Greenberg, D. M. *et al.* The song is you: Preferences for musical attribute dimensions reflect personality. *Soc. Psychol. Personal. Sci.* **7**, 597–605 (2016).
52. Fricke, K. R., Greenberg, D. M., Rentfrow, P. J. & Herzberg, P. Y. Measuring musical preferences from listening behavior: Data from one million people and 200,000 songs. *Psychol. Music* **49**, 371–381 (2021).
53. Flannery, M. B. & Woolhouse, M. H. Musical preference: Role of personality and music-related acoustic features. *Music Sci.* **4**, 205920432110140 (2021).
54. Anglada-Tort, M., Harrison, P. M. C., Lee, H. & Jacoby, N. Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution. *Curr. Biol.* **33**, 1472–1486.e12 (2023).
55. Goupil, L., Ponsot, E., Richardson, D., Reyes, G. & Aucouturier, J.-J. Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nat. Commun.* **12**, 861 (2021).
56. McAleer, P., Todorov, A. & Belin, P. How do you say 'hello'? Personality impressions from brief novel voices. *PLoS ONE* **9**, e90779 (2014).
57. Scherer, K. R. Personality inference from voice quality: The loud voice of extroversion. *Eur. J. Soc. Psychol.* **8**, 467–487 (1978).
58. Banse, R. & Scherer, K. R. Acoustic profiles in vocal emotion expression. *J. Personal. Soc. Psychol.* **70**, 614–636 (1996).
59. Jacobsen, T. & Beudt, S. Domain generality and domain specificity in aesthetic appreciation. *New Ideas Psychol.* **47**, 97–102 (2017).
60. Valentova, J. V. *et al.* Vocal parameters of speech and singing covary and are related to vocal attractiveness, body measures, and sociosexuality: A cross-cultural study. *Front. Psychol.* **10**, 2029 (2019).
61. Henrich, J., Heine, S. J. & Norenzayan, A. The weirdest people in the world?. *Behav. Brain Sci.* **33**, 61–83 (2010).
62. Larrouy-Maestri, P. & Morsomme, D. Criteria and tools for objectively analysing the vocal accuracy of a popular song. *Logoped. Phoniatr. Vocol.* **39**, 11–18 (2014).
63. Larrouy-Maestri, P., Magis, D. & Morsomme, D. Effects of melody and technique on acoustical and musical features of Western operatic singing voices. *J. Voice* **28**, 332–340 (2014).
64. Boersma, P. Praat, a system for doing phonetics by computer. *Glott Int.* **5**, 341–347 (2001).
65. Shue, Y. L., Keating, P., Vicenik, C. & Yu, K. VoiceSauce: A program for voice analysis. In *Proc. ICPHS XVII 1846–1849* (2011).
66. Gordon, M. & Ladefoged, P. Phonation types: A cross-linguistic overview. *J. Phon.* **29**, 383–406 (2001).
67. Yumoto, E., Gould, W. J. & Baer, T. Harmonics-to-noise ratio as an index of the degree of hoarseness. *J. Acoust. Soc. Am.* **71**, 1544–1550 (1982).
68. Murton, O., Hillman, R. & Mehta, D. Cepstral peak prominence values for clinical voice evaluation. *Am. J. Speech Lang. Pathol.* **29**, 1596–1607 (2020).
69. Larrouy-Maestri, P. & Morsomme, D. The effects of stress on singing voice accuracy. *J. Voice* **28**, 52–58 (2014).
70. Lartillot, O., Toiviainen, P. & Eerola, T. A Matlab toolbox for music information retrieval. In *Data Analysis, Machine Learning and Applications* (eds. Preisach, C., Burkhardt, H., Schmidt-Thieme, L. & Decker, R.) 261–268 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008). https://doi.org/10.1007/978-3-540-78246-9_31
71. Bogdanov, D. *et al.* Essentia: An audio analysis library for music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference* 493–498 (2013).
72. Harrison, P. M. C. *et al.* Gibbs sampling with people. *Adv. Neural Inf. Process. Syst.* **33**, 10659–10671 (2020).
73. Woods, K. J. P., Siegel, M. H., Traer, J. & McDermott, J. H. Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psychophys.* **79**, 2064–2072 (2017).
74. R Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, 2021).
75. RStudio Team. RStudio: Integrated development environment for R. In (RStudio, PBC, 2022).
76. Krippendorff, K. Estimating the reliability, systematic error and random error of interval data. *Educ. Psychol. Meas.* **30**, 61–70 (1970).
77. Krippendorff, K. Computing Krippendorff's Alpha-Reliability. (2011).
78. Gamer, M., Lemon, J., Fellows, I. & Singh, P. Various coefficients of interrater reliability and agreement. (2019).
79. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, (2015).
80. Larrouy-Maestri, P. "I know it when I hear it": On listeners' perception of mistuning. *Music Sci.* **1**, 205920431878458 (2018).
81. Finger, H., Goeke, C., Standvoß, K. & König, P. LabVanced: A unified JavaScript framework for online studies. In *International Conference on Computational Social Science* (Cologne, 2017).
82. Fricke, K. R. & Herzberg, P. Y. Personality and self-reported preference for music genres and attributes in a German-speaking sample. *J. Res. Personal.* **68**, 114–123 (2017).
83. Revelle, W. psych: Procedures for personality and psychological research. (2021).
84. Corey, D. M., Dunlap, W. P. & Burke, M. J. Averaging correlations: Expected values and bias in combined Pearson r s and Fisher's z transformations. *J. Gen. Psychol.* **125**, 245–261 (1998).

Acknowledgements

We are thankful to Nori Jacoby for implementing and deploying Experiment 1 in PsyNet, (<https://www.psynet.dev>), a Python package for performing complex online behavioral experiments at large scale. PsyNet is based on the Dallinger framework (<https://dallinger.readthedocs.io>) for hosting and deploying experiments and was developed by the Computational Auditory Perception Research Group at the Max Planck Institute for Empirical Aesthetics. We are also thankful to Zofia Holubowska for assistance in data collection in Experiment 2; to Klaus Frieler for fruitful discussions about data analyses; to Susan Rogers for assistance with the recording of singers at Berklee College of Music; and to Andres Abello for technical support during the recording sessions.

Author contributions

C.B., P.L.M. and D.P. conceived the study. P.L.M. recorded singers. C.B. processed sound files, prepared stimuli for experiments, ran acoustic analyses, handled experimental data collection (with institutional assistance), carried out the statistical analyses and data visualizations and wrote the first draft of the manuscript. P.L.M. supervised all mentioned stages. P.L.M. and D.P. revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-58924-9>.

Correspondence and requests for materials should be addressed to C.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024