Current Psychology
https://doi.org/10.1007/s12144-022-02734-7

Auditory perceptual assessment of voices: Examining perceptual ratings as a function of voice experience

Julia Merrill[1,2]

Accepted: 16 January 2022

Abstract
Understanding voice usage is vital to our understanding of human interaction. What is known about the auditory perceptual evaluation of voices comes mainly from studies of voice professionals, who evaluate operatic/lyrical singing in specific contexts. This is surprising as recordings of singing voices from different musical styles are an omnipresent phenomenon, evoking reactions in listeners with various levels of expertise. Understanding how untrained listeners perceive and describe voices will open up new research possibilities and enhance vocal communication between listeners. Here three studies with a mixed-methods approach aimed at: (1) evaluating the ability of untrained listeners to describe voices, and (2) determining what auditory features were most salient in participants' discrimination of voices. In an interview ($N=20$) and a questionnaire study ($N=48$), free voice descriptions by untrained listeners of 23 singing voices primarily from popular music were compared with terms used by voice professionals, revealing that participants were able to describe voices using vocal characteristics from essential categories indicating sound quality, pitch changes, articulation, and variability in expression. Nine items were derived and used in an online survey for the evaluation of six voices by trained and untrained listeners in a German ($N=216$) and an English ($N=50$) sample, revealing that neither language nor expertise affected the assessment of the singers. A discriminant analysis showed that roughness and tension were important features for voice discrimination. The measurement of vocal expression created in the current study will be informative for studying voice perception and evaluation more generally.

Keywords Voices · Vocal features · Singing · Expression · Phonetics · Speech perception

Introduction

Recordings of singing voices are an everyday, omnipresent phenomenon accessible via portable devices and transmitted in many everyday situations, whether one is driving a car and listening to the radio, going to the store, or watching television or YouTube. Listeners are thus highly familiar with different kinds of musical styles and voices of different expressions. Furthermore, voices are able to express emotions (see Scherer, 1995, on vocal emotion expression in the context of music) and have an impact on the aesthetic judgments of music (see Ackermann, 2019; Ackermann & Merrill, 2021; Greasley et al., 2013; Merrill & Ackermann, 2020, in the context of disliked music). Nevertheless, the evaluation and description of voices has been restricted to trained listeners and to specific contexts. Consequently, what we know about vocal expression from auditory descriptions comes almost solely from studies with trained listeners; and right now, for future research, only those terms and descriptions used in a professional context are available to evaluate judgements and impressions of voices. This is surprising, as little is known about how typical listeners (who not only outnumber the professionals but who are also the targeted listeners/consumers) perceive voices, i.e., how they describe them and what is conspicuous to them. Researchers who have created tools for the description of vocal features in singing voices have already stated that this research should be extended to untrained listeners (Garnier et al., 2007), and to voices other than operatic/lyrical ones (Oates et al., 2006). This will be necessary in order to generalize the findings and actually pursue research with untrained listeners in various

✉ Julia Merrill
julia.merrill@ae.mpg.de

[1] Max Planck Institute for Empirical Aesthetics, Grüneburgweg 14, 60322 Frankfurt am Main, Germany

[2] Institute of Music, University of Kassel, Kassel, Germany

Published online: 27 January 2022

contexts. A tool that untrained listeners can use to describe voices would also enhance research possibilities as well as facilitate the communication about voices between untrained and expert listeners. The need for the latter has already been addressed by focusing on a common terminology between experts from different fields (e.g., Henrich et al., 2008), but not in comparison to untrained listeners. Hence, research is needed toward an understanding of auditory descriptions of voices by untrained listeners, including whether listeners come to similar evaluations between each other and compared to professionals. The goal of the current series of studies was therefore to find free descriptions of untrained listeners and evaluate their ability to use 'professional' terms and find a set of items that would lead to comprehensive vocal profiles of singing voices from different styles.

A "comprehensive" vocal profile can be achieved by evaluating a wide range of vocal features describing the vocal–articulatory expression of a voice, as has been shown by extensive tools for describing the characteristics and particularities of speaking and singing voices with multiple features (e.g., Bänziger et al., 2014; Bose, 2001; Henrich et al., 2008; Laver, 1980; Mathieson, 2001; Wapnick & Ekholm, 1997). The categories of voice description often reference pitch (habitual or average pitch, pitch range, etc.), dynamics (loudness), articulation (intelligibility, sound duration, accentuation, etc.), sound quality (noise, resonance, onsets, etc.), and the variability of all these features (e.g., changes in pitch, loudness, or sounds).

Depending on their purpose, these tools involved compiling items from these categories into an inventory focused most often on the sound quality, e.g., in clinical research on aspects of roughness, breathiness, and hoarseness (for the RBH scale, see Nawka et al., 1994; for GRBAS, see Hirano, 1989) and in Western lyrical/operatic singing on timbre (color/warmth, bright, light, dark, etc.), resonance (ring, full, dull), vibrato, clarity/focus, nasality (nasal, twang) (Ekholm et al., 1998; Garnier et al., 2007; Henrich et al., 2008; Oates et al., 2006). Particularly in singing, the purpose of these tools has been to assess voices in a pedagogical context, evaluating singing technique in Western lyrical or operatic singing, with judges being voice professionals (teachers, singers, etc.).

However, studies using a qualitative approach to evaluating singing voice note that the descriptions provided by singing teachers were not limited to voice quality, but also consisted of references to voice production (articulation, breathing control, etc.), vocal health, the music (musical style, performance, etc.), value judgments (aesthetic judgments, technical level, [dis]agreement, etc.), emotional affect, or the singer's personality (see, e.g., Garnier et al., 2007; Mitchell, 2014). These studies indicate that such criteria merge together to form a "global judgment of value and agreement" (Garnier et al., 2007, p. 73) and that the hedonistic judgment and appraisal of a singer and his or her performance are often foregrounded in the judges' discourse and "an introspection is then required to make explicit all the different criteria which are responsible for this global feeling" (Ekholm et al., 1998; Garnier et al., 2007, p. 74). Likewise, Mitchell (2014, p. 198) notes that teachers even "generally avoid describing the overall sound of the singer."

This notion is of particular interest for research into the psychology of speech, where the focus is on relations between voice perception and evaluations and judgments, e.g., of which vocal features relate (acoustically and perceptually) to emotions (Banse & Scherer, 1996; Scherer, 1995), intentions (Hellbernd & Sammler, 2016), and aesthetic judgments, such as the ideal voice (Hollien, 2000) attractive voice (Babel et al., 2014; Zuckerman & Miyake, 1993), or (dis)liked voice in the context of music (Ackermann & Merrill, 2021; Greasley et al., 2013; Merrill & Ackermann, 2020). Here, judgments by untrained listeners can be helpful for extending research to a larger group of participants, but untrained listeners can use only a few of the existing assessments.

The RBH scale, for example, was developed in the clinical context, and has been used by untrained listeners after training; but it is limited to the perception of noise in the voice (Anders et al., 1988). Another, quite comprehensive tool has been developed with the specific purpose of describing emotional speech; this is the Geneva Voice Perception Scale (GVPS, Bänziger et al., 2014). With this instrument, listeners without any prior training have been able to rate voices according to their loudness, pitch, intonation, sharpness, articulation, roughness, instability, and speaking rate, leading to characteristic descriptions of emotional expression. Hence, it can be assumed that untrained listeners are able to evaluate voices if suitable scales for doing so are available to them. Likewise it can be assumed that exposure to different singing styles constitutes an implicit learning process, which serves as a baseline for such evaluations, as has been shown for related processes, such as assessing correctness in singing (Larrouy-Maestri, 2018) and gaining musical capacities without explicit training (Bigand & Poulin-Charronnat, 2006).

It seems therefore promising to engage closer with descriptions of voices by untrained listeners to find out which terms they use and whether similarities exist between their descriptions and 'professional' descriptions. Ultimately, a set of items assessing the vocal–articulatory expression of singing voices that listeners with differing expertise agree on could be used to augment a listener's perception of voices and to help them express that perception. These vocal profiles can, for example, be used to help explain value judgments of voices (i.e., which features evoke liking/disliking or certain emotions) and to depict untrained listeners' general impressions of voices, but also

to distinguish perceived features from subjective aesthetic judgments (see, e.g., Mitchell, 2014). Hence, by focusing on the auditory assessment, direct connections between the person's perception and the evaluation can be made.

Besides new research possibilities in empirical aesthetics and emotion research, vocal profiles can facilitate the communication between voice professionals and untrained listeners such as voice care clinicians (from phoniatrics, logopedics etc.) and patients, but also between voice pedagogues and students. That there is a need for tools to describe vocal features has been observed in previous research. So far, tools have mainly been developed that bridge the terminology gap between expert listeners from different fields (Ekholm et al., 1998; Henrich et al., 2008). Nevertheless, in contrast to untrained listeners, trained listeners already have an active vocabulary to communicate about voices, which untrained people most likely do not have. Therefore, it is necessary to investigate the verbal abilities of untrained listeners to describe voices. Garnier et al. (2007), who investigated voice quality in lyrical singing, already noted the limitation of only focusing on expert listeners and asked whether non-expert listeners would use the same lexicon to describe voice quality. Oates et al. (2006), who also developed a tool for the description of operatic singing, noted the limitation that the research was confined to just one type of classical singing and asked whether this research could have equal application to other singing styles.

Hence, if untrained listeners are enabled to produce vocal profiles, this will facilitate the perceptual and verbal descriptions of singing voices, which opens up new research possibilities. If those can be applied equally well by expert listeners, a consensual terminology can be achieved that allows for communication between trained and untrained listeners.

Why Research the Singing Voice in Popular Music?

For a long time, musicology and psychology focused on Western classical music and lyrical singing. But vocal expressions with low acceptance in "classical" singing, such as pitch breaks, roughness, and breathiness (Hähnel, 2015); twang (see e.g., Sadolin, 2009; Sundberg & Thalén, 2010); shouting/belting (Stone et al., 2003); pressed phonation (Thalén & Sundberg, 2001); and certain energy distributions in the sound (Cleveland et al., 2001), etc., have been formative for popular music styles. This vast spectrum of vocal expressions warrants investigating a much broader range of features than the sanctioned and standardized styles of classical singing. Also, phenomena of everyday communication have a far greater incidence in popular music than in classical singing; these include crooning (whispering) and moaning (whining, lamenting, and sighing), and howling or wailing. Furthermore, both the emotions expressed and the features employed in singing popular music are much closer to those found in everyday life, in speaking (i.e., they approximate speaking more than they do opera) and in 'everyday' singing. Hence, using popular music singing styles, the listener has the chance to describe a variety of vocal features, which are typically more prominent than in speaking voices, but still similar to (heightened) speech. Investigating voices in this rich environment considers emotional reactions and the influence of personal attitudes on the perception of vocal utterances and therefore serves as a basis for further research on the impact of popular music on our lives and the communicative role of the singing voice in music.

Reliability and Agreement in Voice Evaluation

It is evident that inventories with a small number of items from a single category (e.g., sound quality or noise) lead to higher reliability than others that have up to 20 items from different categories (Webb et al., 2004). For the evaluation of singing voices, higher reliability measures (i.e., the intra–class correlation coefficient, or ICC) have been found for sets of items that focus on sound quality. Ekholm et al. (1998) had seven voice teachers evaluate 16 singers of Western lyrical singing (with piano accompaniment) on four sound qualities (color/warmth, resonance/ring, appropriate vibrato, clarity/focus), thus revealing an interrater reliability of 0.44. Likewise, Oates et al. (2006) investigated a set of items for rating operatic singing (appropriate vibrato, ring, pitch accuracy, evenness throughout the range, and strain) using nine trained judges and 21 voices, and showed a high interrater reliability of 0.75. A wider set of items was evaluated by Wapnick and Ekholm (1997), who report twelve perceptual criteria on three factors—intrinsic quality (color/warmth, dynamic range, etc.), execution (flexibility, intonation accuracy, etc.), and diction (single item)—that were rated by 21 voice teachers evaluating 21 voice samples, with a mean interrater reliability of 0.49.

Other, more extensive, studies on Western lyrical singing have not reported (inferential) statistical comparisons. Garnier et al. (2007) interviewed 11 voice teachers evaluating 18 musical pieces by three singers, and report consensus of verbal descriptions of vocal quality, that is nasal, full, bright, dull, light, dark, presence of vibrato, and items on placement of the voice. The most extensive tool for evaluating Western lyrical singing was developed by a multidisciplinary group around Henrich et al. (2008). The "listening sheet" they developed covers not only aspects of breathing, vibrato, and placement, but also phonetic aspects (segmental and suprasegmental level, vowels and consonants, phrase, accents, intelligibility), sound color (high/low pitch, ring, energy spectral distribution, dark/light), sound intensity, and pitch (powerful/weak, efficiency, vocal effort, singing formant, voice range). The validation of the tool was done descriptively with (semi–)professionals trained ($N = 6$) and

untrained ($N = 18$) with the listening sheet. The GVPS (Bänziger et al., 2014) reported ICCs between 0.216 and 0.811 per feature, which shows that untrained listeners ($N = 19$, evaluating 160 emotion portrayals) were able to judge a set of items of various categories with good agreement. While the GVPS used at least a large number of stimuli, the number of judges was in all investigations small.

Nevertheless, interrater reliability is considered problematic in the auditory vocal assessment. First, while methods such as an ICC inform about agreement using several voice samples, they do not reflect variations of agreement for specific samples (Kreiman & Gerratt, 1998); also, interrater variability might be an issue of task design, not of listener unreliability because, among other things, it is dependent on the magnitude of the attribute being measured in a voice (Kreiman et al., 2007). Therefore, other matching procedures need to be applied.

Another aspect to consider in voice assessments is the use of bipolar scales. Bipolarity is an issue in voice description as some features can be well represented with opposing poles such as high–low or dark–bright, but features such as rough or breathy were better described with "applicable–not applicable" (Henrich et al., 2008). A bipolarity, however, is often helpful for untrained listeners as one pole can explain the other, e.g., rough–smooth. Methodologically, bipolar rating scales can conflict with conventional approaches on psychological test construction because a normal distribution might not be achieved. For example, if an item can be well applied to a voice, ratings should be skewed, because raters univocally decide to one direction of the ratings scale; in the case of many voices, the distribution can be bimodal (e.g., if voices can be assigned to either high or low pitch).

The Present Study

In a mixed-methods approach, free descriptions of voices were collected in interview sessions and analyzed as to the terms participants use and whether these cover important categories of (professional) voice descriptions. Second, these untrained listeners were approached with 'professional' terms from different fields to see which ones they were able to use properly. The free descriptions and professional terms were combined into a list of features to describe vocal expression and investigated on the usage and agreement among larger groups of participants. The final online survey combined participants having various backgrounds with voices (self-reported on the speaking and singing voice), which enabled the investigation of differences in voice assessment as a function of expertise. As the studies were conducted in German, the terms were made applicable to the English speaking research community by translating and separately evaluating the set of features with English native speakers.

Study 1: Interview Study

In an interview study, participants expressed voice descriptions in the context of disliked voices. It is in precisely this area that potential can be seen for causing listeners to formulate differentiated descriptions, because negative emotions lead to stronger cognitive processes and have stronger effects on behavior, and there are more words for expressing negative than for positive feelings (Baumeister et al., 2001; Rozin & Royzman, 2001). Likewise in voices studies, the evaluations were biased by appraisal, and judges seemed to be more able to describe less appealing voices and less able to articulate the sound quality of their preferred voices, that is, "the more beautiful the voice is perceived, the more difficult it is to describe" (Mitchell, 2014, p. 194). The interviews were accompanied by a list of terms used by professionals to describe vocal expression in order to check the terms' usability by the untrained participants. The approach to the data was a qualitative one, including a descriptive level of comparison to voice professionals. Inferential statistics were possible in the follow-up studies.

Methods

Participants

Twenty participants (15 female, 5 male) were recruited (convenience sample), each of whom brought one or two recordings of a disliked (professional) singer to the interview session, with 23 voices being incorporated into the study (Table S1, supplementary material). Participants were on average 38.55 ($SD = 16.39$) years old, and 13 were university students. The chosen music titles represented mainly popular music styles; nine singers were female, 13 male, and one was a child (Angelo Kelly); and at least two singers were classically trained (Sarah Brightman and Anneliese Rothenberger). All experimental procedures were approved by the Ethics Council of the Max Planck Society, and were undertaken with written informed consent of each participant.

Procedure

The session consisted of two parts. In the first part, after listening together to the full song, the interviewer asked the participant to describe the voice; in the second part, the participant was asked to select, on a questionnaire, vocal characteristics that described the disliked vocal features. Each session lasted about an hour, and monetary compensation was 10 euros.

The questionnaire consisted of an adapted and shortened version of the Catalogue of Vocal-Articulatory

Expression (Bose, 2001), and comprised items from the major categories of voice evaluation. The questionnaire had already been used with trained listeners for the evaluation of "speech song" in 20th–century Western art music (Merrill, 2017; Merrill & Larrouy-Maestri, 2017) and jazz singing (Merrill, 2019). Twenty items, including 16 bipolar items, were evaluated; these were: average pitch (low–high), timbre (dark–bright), loudness (loud–soft or "quiet," being closer to the German original), tension (tense–unsupported or "un-tense"), sonority (full–thin), faucal distance (wide–constricted or "tight"), sound of voice (soft–hard), resonance (dull–shrill), sound duration (lengthened–shortened), pitch changes (sudden–gliding), mode of phonation (speaking–shouting; note a different use of this expression in Sundberg, 1975), vibrato (vibrato, tremolo), onsets (soft–hard), articulation precision (precise–imprecise), sound modulations (little–much), pitch modulations (little–much), and breathy, rough, creaky, and nasal.

In addition to the untrained listeners, three scientists, each holding a doctorate (Ph.D.) in speech science and experienced in the assessment of singing voices, also evaluated the 23 voices (Table S2). Three was considered a sufficient number of voice professionals for the current purpose, i.e., to provide a descriptive level of comparison with the evaluations by untrained listeners. Note that Wapnick and Ekholm (1997), who used quite an extensive questionnaire for voice evaluation, found an interrater reliability of $r = .75$ with three professional judges.

The analysis included the following steps for each participant: (a) characteristics in the questionnaire were linked to those named in the interview, with analogies sought between both; and (b) the questionnaire data were descriptively compared to the evaluations by the trained listeners. If a participant's assessment was in agreement with two of the trained listeners, the characteristic was considered to be assessed "correctly" (i.e., showed a match).

Results

It should be noted that results are based on German terms and have been translated into English. Most verbal descriptions refer to the sound of the voice ($n = 29$), such as nasal, miffed, pressed, knödel (i.e., retracted), less soft, thin, pointy, squeaky, squealing, "cheeping," tinny, and "without substance." Some terms fit the category of tension, such as "pressed" and "squeaky," but were used by the participants based on the sound and not on a functional understanding of muscle tension. Of these sound features, some refer to noise ($n = 8$), such as rough, scratchy, noise in tone onset, scrabble/paw, not clear, "unclean" (in the sense of noise), which showed that the terms indicated a perceived roughness. Other terms such as "pressed" and "scrabble" were also used in part to describe noise ($n = 6$). Other aspects of tension refer to the impression of a powerless and weak voice.

There were ten descriptions of average pitch, such as a changing or "wrong" pitch level, "unclean" (in the sense of intonation or the sense of pitch changes), high, low, monotone, missing depth, or of singing/reaching tones from below. A total of 13 statements refer to the overall impression of the voices, including seven descriptions of sounding childlike, not masculine, arrogant, unappealing, disinterested ("uncouth," "dashed off"), boring, whiny/sniveling, moaning, eerie, and exaggeratedly female. The participants only rarely mentioned features concerning articulation, accents, intelligibility, sound duration, mode of phonation, and rhythmic aspects.

Regarding the questionnaire, the number of ratings are depicted in Fig. 1 and are shown in comparison to those used by the trained listeners. Notably, "creaky" and "breathy" are both rarely mentioned and seldom used correctly. Although nasal is used more frequently, it is only used correctly in less than 50% of cases. Sonority and sound of voice are used correctly with just over 50%, modulations of tones, and faucal distance with just over 60%. Some features are rarely

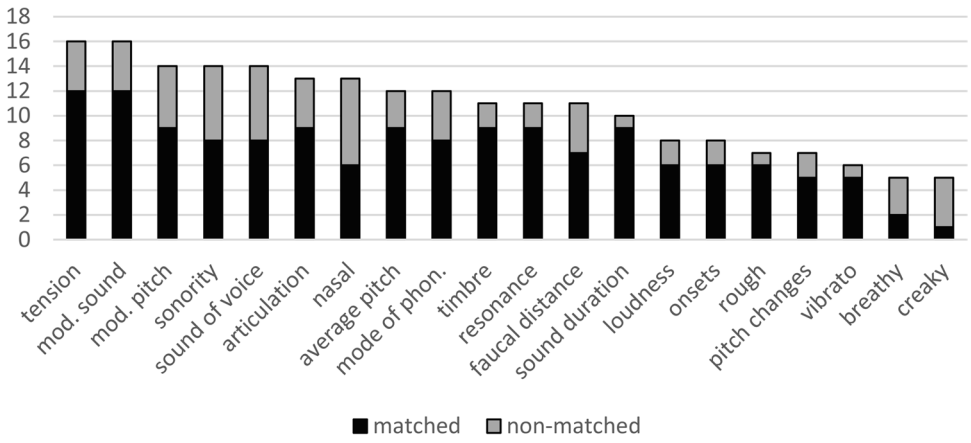

Fig. 1 Frequencies of the Disliked Characteristics in the Questionnaire of Study 1. *Note*. Number (y-axis) of matched and non-matched characteristics (x-axis). Participants only checked the disliked features of the 23 singers

mentioned; hence, a "correct" use should be considered with caution, and results need to be contextualized with the verbal descriptions provided by the participants.

Interim Discussion

The participants utter their impressions of voices on different levels that mirror major categories of vocal description (e.g., Bose, 2001; Wapnick & Ekholm, 1997). The utterances also include metaphors and associations, and the participants mix aesthetic judgments into their descriptions of vocal features (Garnier et al., 2007; Mitchell, 2014). The vocabulary was certainly limited but on average, each participant was able to name two adjectives that described the particularities of a singer's voice.

Considering the terms expressed by the participants and the overlap with those they chose from the questionnaire, conclusions can be drawn on the usage and application of the terms. In the case of ambiguities, combinations or changes of items are discussed for the follow-up study. Overall, some features could be well applied, i.e., they mostly matched the ratings of the trained listeners and reflected the verbal expressions of the participants. These include mode of phonation (speaking–shouting), sound of voice (hard–soft; the specification "onset" was removed as it was redundant), average pitch (low–high) and timbre (dark–bright), which were already addressed by the participants in the interviews and therefore belonged to the active vocabulary of the listeners. The characteristics of the feature pitch changes (sudden–continuous) were used frequently in the interviews and on the questionnaire.

The feature modulations of sound and pitch showed considerable overlap, i.e., the participants did not differentiate between the two assessments (and were therefore combined in the next study). The feature tension could be well applied, but as the participants used the term "pressed" in their utterances more often than "tense," another item was created with the poles "lax–pressed" to investigate the overlap between both in the next step. The features sonority and loudness showed ambiguous usage and need further evaluation in order to come to better conclusions. The feature "roughness" was the best applied of the all noise features and was frequently mentioned. Other features describing noise, such as creaky and breathy, were (also after further inquiry) unknown to participants or confused with roughness, which had already been shown in other studies investigating the distinction between breathiness and roughness (Anders et al., 1988; Kreiman & Gerratt, 1998).

The feature articulation described only articulation precision and fell short of being able to help express particularities in articulation. As the latter turned out to be an important feature for the participants, the feature was changed in order to focus on overall (and rather broad) peculiarities of articulation. The features faucal distance, nasal, and resonance (dull–shrill) showed considerable overlap. In the questionnaire, the participants mixed faucal constriction and perceived nasality with descriptions from the interviews such as "squeaky," which suggests that for the participants the terms are connected or cannot be differentiated. In cases where the participant selected "nasal," trained listeners tended to perceive faucal constriction, which suggests a combination of the characteristics squeaky and nasal. The characteristic "wide," however, remained applicable and could be combined with the resonance characteristic "dull" inasmuch as its opposite, "shrill," was linked with "squeaky."

Other features, such as sound duration (lengthened–shortened), were used somewhat diffusely in the interviews and questionnaire and did not seem to be understood or did not play a role in the description of voices. Vibrato ($n = 2$) and trembling ($n = 4$) were not mentioned in the interview and hardly in the questionnaire. The participants' comments suggested that they were not familiar with the term vibrato, so the usage seems to depend on their expertise and interest, and is probably more relevant in the context of classical singing. Non-singers, for example, are very divided on ratings of vibrato, and their assessment of a performance might not depend on vibrato, unlike singers' evaluations (Reddy & Subramanian, 2015).

Study 2: In-House Questionnaire Study

Study 2 is to be understood as an intermediate step toward the online survey. The newly created features based on Study 1 were evaluated with more participants but still in a controlled setup in an in-house study. Participants evaluated the voices from the interview study using the new feature list and provided ratings on the difficulties of the items as well as free text to explain these issues. The ratings were again descriptively compared to the expert ratings from Study 1.

Methods

Participants

Forty–eight participants (29 female, 19 male) with a mean age of 40.06 years ($SD = 18.24$) took part in this study (convenience sample). Twenty–one had at least a high school degree, 27 were university students. Five participants were professionally involved in music, 26 played an instrument.

Questionnaire

The set of features consisted of 13 bipolar items: average pitch (low–high), loudness (loud–soft or "quiet"), sonority (full–thin), timbre (dark–bright), sound (soft–hard),

tension (tense–unsupported or "un-tense"), noise (rough/scratchy–smooth), resonance (dull/wide–squeaky/nasal), pressed (lax–pressed), pitch changes (sudden–gliding), articulation (distinctive–plain), mode of phonation (speaking–shouting), expression (varied–uniform).

Procedure

Sixteen of the 20 titles of the interview study were selected (Table S3) and a representative excerpt was chosen from each song and a balanced ratio of male and female singers was ensured. Features were evaluated on a 4–point scale, which did not leave a midpoint, so the participants had to decide between the characteristics (comparable to forced choice). Additionally, liking was assessed for each feature separately as was the general liking of voice, song, text, and musical style (each on a 5–point scale; not part of the current report) and the familiarity of the singer (yes/no; name of the singer).

At the end of the study, participants were asked how well they were able to rate their impression of the voice with the given characteristics (on a 5–point scale, from very good to not good at all) and if there were characteristics that seemed to be missing (no, yes, namely...). Finally, participants were asked to evaluate each feature on its difficulty (4–point scale from "not at all" to "very much") and to justify the rating in an open comment field.

Each participant sat in a booth in a group testing room and rated on six out of a total of 16 singers, with one singer rated by all participants, The Tallest Man on Earth (no. 16). The music samples were listened to via headphones (Beyerdynamic DT 770 Pro), whereby the volume could be individually adjusted by the participant. The music titles were played back according to lists created beforehand so that (a) the pieces were evaluated as evenly as possible throughout the study, and (b) the order in which they were presented varied. The questionnaires were filled out on paper. Altogether, each singer was rated by 15 to 18 participants. The session lasted approximately 1–1.5 h, and each participant received monetary compensation of 15 euros. Data were analyzed using SPSS 25.

Results

Interrater Agreement

In order to compare interrater reliability with other studies investigating listener agreement in voice evaluations, an intra–class correlation (ICC; model: two–way random effects; type: absolute agreement) was performed for the whole questionnaire (13 items) for the one singer all participants rated (no. 16). The ICC revealed an average measure of .212 with a 95% confidence interval from −.073 to .473 ($F(45.540) = 1.354$, $p = .067$). Four features showed (Table S4) negative corrected item–total correlation, that is speaking–shouting (−.383), rough/scratchy–smooth (−.156), tense–unsupported (−.094) and lax–pressed (−.045). Hence, the average ICC is lower than in comparable studies where it was around $r = .4$ (Bänziger et al., 2014; Merrill & Larrouy-Maestri, 2017; Wapnick & Ekholm, 1997; note that different methods were used to calculate these correlation coefficients).

Application of Features

As mentioned above, the ICC does not allow for conclusions about feature ratings in individual singers. In order to investigate whether participants chose the same direction of an evaluation for the features in each singer, observed and expected frequencies adapted to a normal distribution were compared with chi-square tests separately for each voice and feature (Table S5). A normal distribution would be expected if the participants could not evaluate a specific feature in a voice well (similarly to a semantic differential scale); a consistent evaluation would be in either direction of the rating scale. This non–parametric statistic was chosen because of the short, ordinal rating scale. After the distributions were inspected visually to make sure that the ratings clearly pointed in one direction, all significant features were compared with the binary ratings by the trained listeners in order to make a descriptive comparison. The results show that no feature revealed unanimous evaluations for all singers. The maximum of significant chi-square tests was found for loudness (in 12 singers), followed by sound of voice, mode of phonation, pitch changes, average pitch, noise, and timbre. All other features showed unanimous evaluations for less than half of the singers. The average rating per singer was compared descriptively to the (categorical) expert ratings showing that the number of non–matched ratings ranged between one to three singers per feature (Fig. 2).

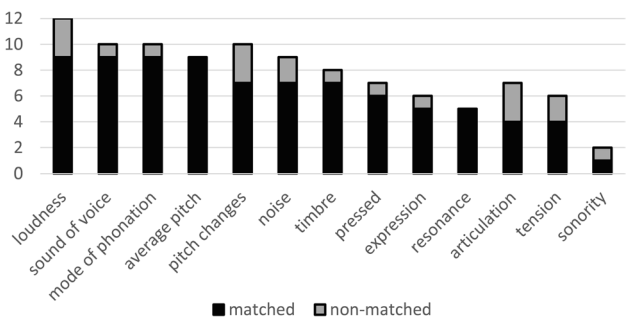

Fig. 2 Number of the Significant Chi-Square Tests in Study 2 Indicating Univocal Ratings. *Note.* Univocal ratings mean ratings into one direction on the bipolar ratings scale. Number (y-axis) of matched and non-matched features (x-axis) compared to the evaluations by the trained listeners from the max. 16 singers

Correlation of Items

Spearman correlations were performed for all items in order to identify possible overlap between features. Higher correlations ($r > .4$; Table 1) were found for average pitch and timbre (e.g., high and bright), loudness and mode of phonation (e.g., loud and shouting), sound of voice and noise (e.g., hard and rough), and sound of voice, tension, and pressed (e.g., hard, pressed, tight).

Difficulties Reported in Evaluation and Self–Assessment

Participants were asked to rate and comment on the perceived difficulty of each item. They reported being able to reproduce their impression of the voice very well ($M = 1.96$, $SD = 0.87$; on a 5–point scale) and did not report any major difficulties with the items (mean ratings between 1.4–2.7 on a 4–point scale; Table S6). Slightly higher rated were dull/wide, tension, mode of phonation, and articulation; smooth, pitch changes, and timbre were slightly lower. The most commented features were "tense–unsupported" (47.9%) and "dull/wide" (43.75%), followed by "speaking–shouting" (33.3%), "low–high" (29.2%), whereby in almost all cases the perceived overlap with "dark–bright" (29.2%) was mentioned as well as on articulation and loudness (27.1% each).

Interim Discussion

The participants did not report major problems with the questionnaire and were able to reproduce their impression of the voices with the given features. Nonetheless, the overall ICC was low, but the items with a negative item-total correlation showed overlap with items perceived to be more difficult, hence, the participants' perception reflected the issues with some features. The comparison to the expert ratings was still only descriptive, but likewise underlined the difficulties reported. The application of the features revealed that differences in feature usage were also dependent on the voice, i.e., in some voices a feature was more evaluable than in another.

In light of the follow-up study, a new combination of items was to be decided, for which all analysis steps presented were considered together. Some features could be well applied, that is vocal expression and average pitch. Some were combined and needed adjusted poles. That is, first, tension and pressed, where the pole "lax" was replaced by "pressureless"; second, roughness and sound of voice, where the pole "smooth" was criticized; third, resonance and timbre, where the pole "dark/wide" was replaced by "dark/dull." The feature timbre itself seemed to be obsolete because it was confused with average pitch, showing that a high voice was often associated with a bright timbre. Mode of phonation was changed so that the poles represent

Table 1 Correlation coefficients (Spearman) between all features

	Pitch	Loudness	Sonority	Timbre	Sound	Tension	Noise	Resonance	Pressed	Pitch changes	Articulation	Mode of phonation	Expression
Pitch	1.00	−0.16	−0.14	0.61	−0.12	−0.02	0.32	0.20	0.04	−0.08	−0.03	0.15	−0.06
Loudness	−0.16	1.00	0.12	0.01	−0.34	0.31	0.04	−0.19	−0.28	0.35	0.28	−0.47	0.24
Sonority	−0.14	0.12	1.00	−0.05	0.26	−0.07	−0.28	0.19	0.28	−0.18	0.24	−0.07	0.25
Timbre	0.61	0.01	−0.05	1.00	−0.16	0.00	0.34	0.20	−0.03	−0.08	0.06	0.05	−0.05
Sound	−0.12	−0.34	0.26	−0.16	1.00	−0.44	−0.43	0.16	0.51	−0.32	−0.22	0.26	−0.16
Tension	−0.02	0.31	−0.07	0.00	−0.44	1.00	0.21	−0.08	−0.54	0.32	0.16	−0.28	0.17
Noise	0.32	0.04	−0.28	0.34	−0.43	0.21	1.00	−0.05	−0.33	0.17	0.05	−0.06	0.02
Resonance	0.20	−0.19	0.19	0.20	0.16	−0.08	−0.05	1.00	0.25	−0.18	−0.11	0.09	−0.10
Pressed	0.04	−0.28	0.28	−0.03	0.51	−0.54	−0.33	0.25	1.00	−0.26	−0.12	0.27	−0.12
Pitch changes	−0.08	0.35	−0.18	−0.08	−0.32	0.32	0.17	−0.18	−0.26	1.00	0.19	−0.32	0.26
Articulation	−0.03	0.28	0.24	0.06	−0.22	0.16	0.05	−0.11	−0.12	0.19	1.00	−0.21	0.34
Mode of phonation	0.15	−0.47	−0.07	0.05	0.26	−0.28	−0.06	0.09	0.27	−0.32	−0.21	1.00	−0.33
Expression	−0.06	0.24	0.25	−0.05	−0.16	0.17	0.02	−0.10	−0.12	0.26	0.34	−0.33	1.00

singing and speaking, in line with the comments. Depending on the research question, this pole could be changed in future research to shouting, which is a common feature in popular music styles (e.g., gospel music shouting, belting, metal singing). Sonority and loudness were disregarded because of the difficulty of evaluating them in recordings (especially of popular music) due to the recording technique (a comment made by both trained and untrained listeners). With regard to articulation, it did not seem to have been apparent to the participants that particularities of articulation beyond pure intelligibility needed to be evaluated. Therefore, a feature of "precise–imprecise" articulation was added to "peculiar–ordinary" (now reformulated), so that both facets could be evaluated.

Study 3: Online Survey

In this final step, a large group of participants took part in the survey consisting of two language groups and people having different backgrounds with voices, i.e., with speaking and singing voices. So far, in voice research, different terminologies between fields were investigated, but only between experts (e.g., voice care clinicians and vocal pedagogues) and not untrained listeners, which was already mentioned as a research gap (Garnier et al., 2007). Therefore, Study 3 was conducted with trained and untrained listeners in order to see whether the untrained listeners came to different ratings of the singers than the experts. The two language groups were done in order to have a proper translation and evaluation of the German questionnaire.

Methods

Participants

Overall 266 participants (165 female, 100 male, one not stated) took part in the study (convenience sample), of which 216 participants (138 female, 77 male, one not stated) completed the German version and 50 participants (27 female, 23 male) the English version of the study. In the German version, the participants were on average 31.31 years old ($SD = 12.19$; range 18–68). 195 had at least a high school degree, and just over half of the participants were university students ($n = 113$). The open question on professional or non-professional involvement with voice led to identification of three groups: 53.24% ($n = 115$) without voice experience, 19.98% ($n = 41$) with speaking voice experience (e.g., speech scientist, speech therapist, actor), 27.78% ($n = 60$) with singing voice experience (e.g., professional singers, singing lessons during their studies, an extensive choir experience of more than 8 years plus individual lessons of more than 2 years, or with background in classical and popular singing styles).

In the English version, participants were on average 35.28 years old ($SD = 14.17$; range 18–69). All but one participant had at least a high school degree, 18 were university students. Concerning professional involvement with voices, 33 had no experience, only two had experience with speaking voices/phonetics, and 15 with singing voices (mostly classical music).

Questionnaire

The new set comprised nine items, which still represent major categories of vocal articulatory expression: average pitch (low–high), noise (rough/scratchy–soft), tension (pressed–pressureless), timbre (squeaky/nasal–dark/dull), pitch changes (sudden–gliding), articulation precision (precise–imprecise), articulation peculiarities (peculiar–ordinary), mode of phonation (sung–spoken), and vocal expression (varied–uniform).

Selection of Voices

Since this investigation was aimed at the everyday listening experience, singers were again investigated with background music. Participants were instructed to pay attention to the voice and to disregard the music as much as possible.

To keep the procedure short, only six singers were selected for evaluation. The choice of singers was made according to the following criteria: The number of singers should not unduly prolong the survey processing time for the participants. Singers were selected so that the features discussed so far could be compared between the studies. Six popular voices and songs were chosen (Table 2), which were expected to be known by the participants due to their age and positions in the charts. With this, the decision was made to control for familiarity, with the limitation that participants' evaluations were not free from personal memories and experiences.

Three female and three male singers were selected, who can be heard well in the recording (a little additional reverb can be heard in Houston). Each excerpt was about 30 s long and was a representative part of the song, in which the voice did not change too much in sound and technique.

Procedure

After the collection of demographic data, including native language, an open question was asked about the participant's involvement with the singing or speaking voice in a private or professional context. The presentation of the six selected singers was in a randomized order. After listening to one recording completely, evaluations of the

Table 2 Selection of singers for the online survey

Singer	Title	Album	Released	Excerpt
Bob Dylan	Don't Think Twice It's All Right	The Freewheelin' Bob Dylan	1963	00:07–00:43
Elvis Presley	Love Me Tender	Love Letters from Elvis	1971	00:02–00:41
James Brown	I Got You (I Feel Good)	Out of Sight	1964	01:26–02:03
Tina Turner	What's Love Got to Do with It	Private Dancer	1984	00:45–01:12
Wanda Jackson	Let's Have a Party	Wanda Jackson	1960	00:04–00:33
Whitney Houston	One Moment in Time	One Moment in Time: 1988 Summer Olympics Album	1988	00:59–01:30

voice included a question on liking and a list of evoked emotions along with the ratings of the vocal features on a 6-point scale (again to force choices). During all steps, the recording could be listened to repeatedly. After this, the liking of the song was evaluated (in the same way as the liking of the voice) and the familiarity with the song (yes/no). This process was repeated until all six voices were evaluated. After room for comments, it was possible to take part in a raffle in which every tenth participant won an Amazon.de voucher worth 10 euros, in the English version an Amazon.com voucher worth 20 US dollars. The survey lasted about 20 min. (Note that the ratings on emotions and liking are not part of the current report.)

For the English version, the German questionnaire resulting from the online survey was translated by the author based on vocabulary common in literature from speech science, phonetics, and voice clinics. It was then checked by two English native speakers who also spoke German and discussed and adapted. Another online survey was started with the same procedure used for the German version.

Results

Effects of Expertise and Language

A repeated-measures analysis of covariance (ANCOVA) with within-subject factor Feature and between-subject factor Expertise and covariates Singer and Language was performed. Levene's test of equality of error variances only became significant for mode of phonation ($F(2,1593) = 6.246$, $p = .002$). Test of between–subject effects was significant for Singer ($F(1) = 72.063$, $p < .001$), but not for Language ($F(1) = .072$, $p = .789$) nor for Expertise ($F(2) = 2.163$, $p = .115$; see Fig. 3). This result shows that both language versions could be used equally well (German and English) and for both trained and untrained listeners.

Interrater Agreement

An ICC for all nine features revealed an average measure ICC of .392 with a 95% confidence interval from .345 to .436 ($F(1595, 12{,}760) = 1.688$, $p < .001$) (Table S7), hence, very similar to comparable studies (Bänziger et al., 2014; Merrill & Larrouy-Maestri, 2017; Wapnick & Ekholm, 1997).

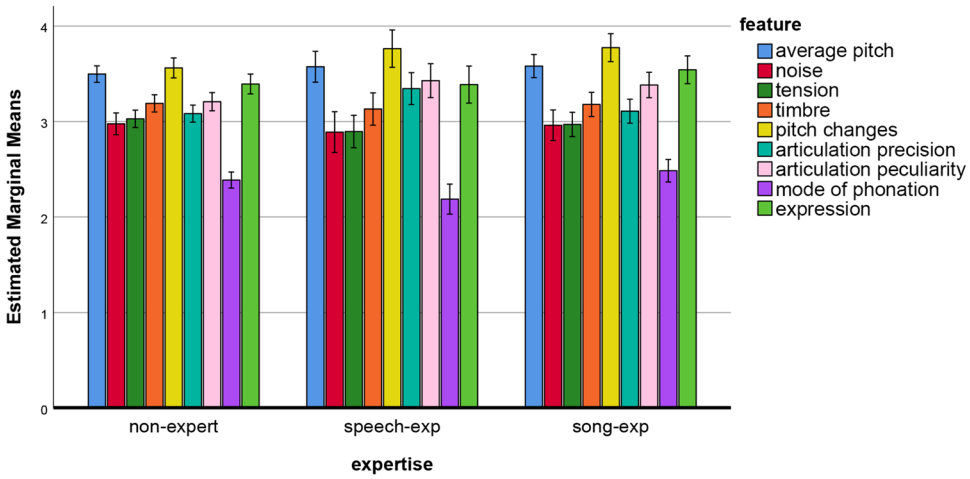

Fig. 3 Effects of Expertise on the Feature Ratings. *Note.* Estimated marginal means for all nine features and the three (non-) expert groups following the ANCOVA. Covariates appearing in the model were evaluated at the following values: singer = 3.50, language = .81. Error bars: 95% confidence intervals

Application of Features

As in Study 2, chi-square tests were used to compare the observed distribution of the features separately for each voice and feature with an expected (normal) distribution over the 6–point rating scale (Fig. 4). All chi-square tests became significant (Table 3).

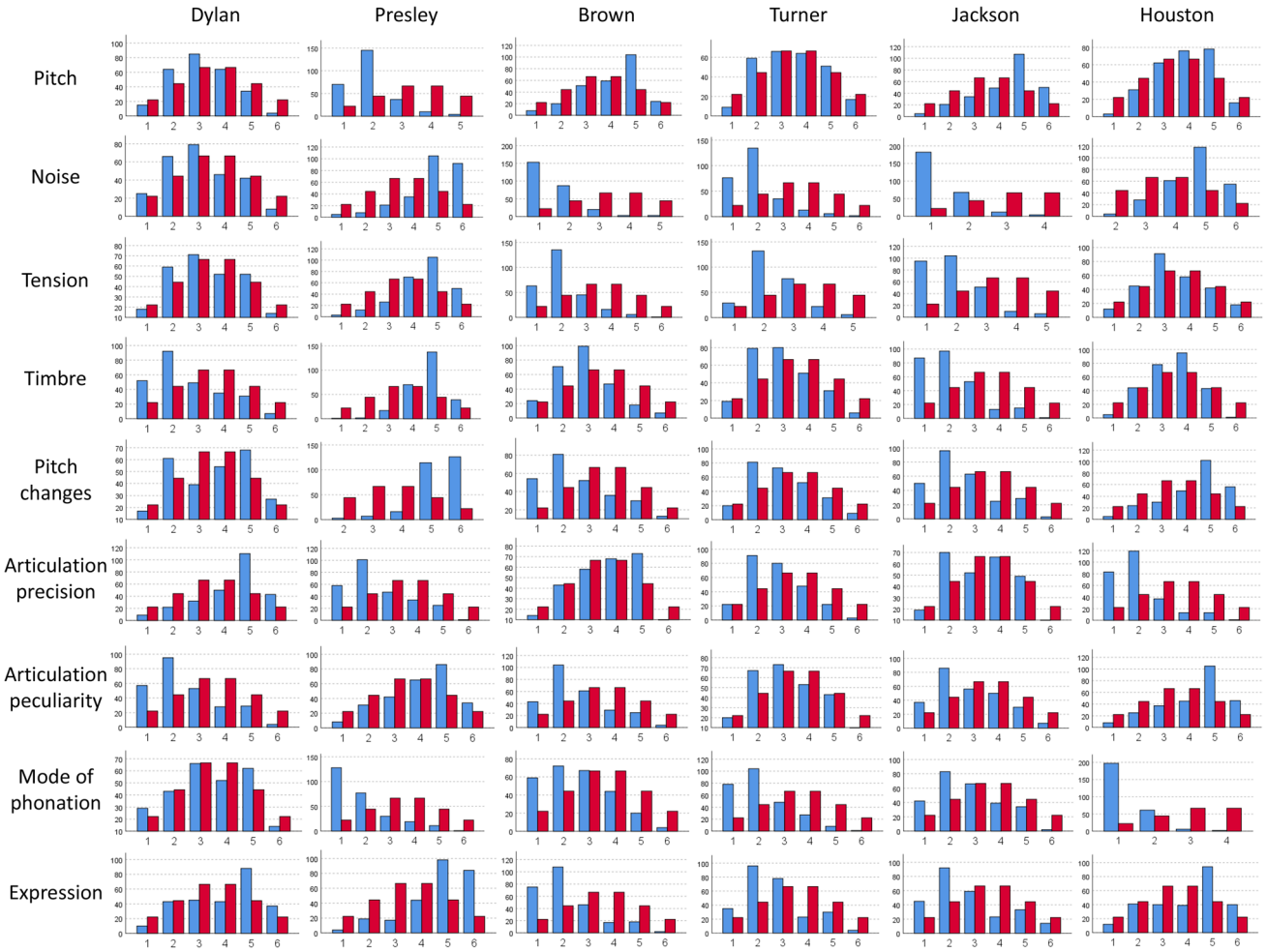

Fig. 4 Observed (Light Blue) vs. Expected (Dark Red) Distributions of Ratings. *Note.* Scale according to characteristics of the items (see text or Table S9). All chi-square tests were significant, but in some cases, small differences were seen: tension, pitch changes, and mode of phonation in Dylan; articulation precision in Brown; tension and timbre in Houston

Table 3 Chi-Square values for singers and features (N = 266, df = 5)

Feature	Singer					
	Dylan	Presley	Brown	Turner	Jackson	Houston
Average Pitch	33.570	451.752	107.323	14.977	169.594	49.526
Noise	28.797	392.143	967.128	421.549	1335.023	252.647
Tension	13.436	182.970	359.301	262.211	426.541	15.692
Timbre	125.316	304.143	63.812	49.714	337.594	47.752
Pitch changes	34.880	748.143	101.617	46.165	143.120	173.910
Articulation precision	157.910	180.602	29.383	84.835	25.647	391.188
Articulation peculiarities	157.865	67.602	144.805	21.895	69.850	146.165
Mode of phonation	15.361	628.602	114.331	299.534	83.602	1569.316
Expression	74.910	311.271	294.511	117.609	109.977	96.835

It was already suspected in Study 2 that the assessment of voices depends not only on the goodness of feature ratings, but also on the explicitness of the vocal feature to be evaluated. It can be seen that chi-square values differ between features and singers and after visual inspection of the distributions (Fig. 4), some evaluations were less clear than others: For example, while for Presley all features were indecisively rated into one direction on the bipolar rating scale, the results for Dylan reveal much lower values for several items and less clear distributions into one direction. A look at the features reveals that while expression has higher values for all singers, tension shows lower values for two singers. Overall, it can be concluded that there were no systematic issues in the evaluation of the characteristics, and each feature could be evaluated well for most of the singers. This means that the evaluations are singer specific and show how unambiguously a feature is expressed; hence they depend less on understanding the characteristics than on the expression of a singer (Kreiman et al., 2007).

Table 4 Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Feature	Function				
	1	2	3	4	5
Explained Variance %	75.8	16.8	5.7	1.3	0.3
Noise	0.823*		0.381		
Tension	0.462*				
Art. precision		0.582*		0.512	
Mode of phonation		0.576*			
Art. peculiarities		−0.342*		−0.331	
Timbre	0.355		−0.701*	0.412	
Average pitch	−0.301	−0.365	0.696*		0.394
Expression	0.359			−0.465	0.475*
Pitch changes	0.436				0.449*

Variables ordered by absolute size of correlation within function. * Largest absolute correlation between each variable and any discriminant function. Loadings < |.3| omitted

Classification of Singers

In order to show which characteristics in which combination contributed best to the differentiation of singers and which features were possibly most salient for the auditory discrimination, the six singers were classified based on the evaluated characteristics using canonical discriminant analysis (CDA). Other studies have used discriminant analyses to assign specific vocal characteristics to different emotions (Bänziger et al., 2014) or intentions (Hellbernd & Sammler, 2016) in speech. Chi-square tests with Wilks' Lambda showed five functions that had a statistically significant effect on the discrimination (Functions 1 to 5, $p \leq .001$).

The first function explained 75.8% of the variance with highest loading items being noise and tension, hence revealing the largest influence. The second function explained 16.8% with articulation precision and peculiarities as well as mode of phonation. The explained variance of the other functions was low (Table 4).

Here, 63.0% of cross-validated grouped cases were correctly classified (63.9% of original grouped cases). Dylan was correctly classified in 66.5% of cross–validated cases, Presley in 82.5%, Brown in 48.9%, Turner in 50.8%, Jackson in 53.8% and Houston in 75.6%. Brown was classified as Jackson in 22.6% of cases, Jackson as Brown in 26.7%. Turner was also confused with Brown and Jackson in 16.5% and 18.4% of cases, respectively (full classification results in Table S8).

Based on the first function, the voices can be distinguished very well by the features noise (rough/scratchy–soft) and tension (pressed–pressureless). Figure 5 shows that Brown and Jackson are positioned in the direction of "rough and pressed" and Presley in the direction of "soft and pressureless." Following Function 2, the voices can be separated based on articulation and mode of phonation, where the poles are determined by Dylan and Houston. However, the distance here is not as large as in Function 1 (also reflected in the lower explained variance).

The functions of the CDA revealed similarities and differences between the singers as well as the features. The fact that noise and tension (Function 1) are similar is shown, for example, by Jackson and Brown, who sing with a rough/scratchy voice as well as with high pressure (pressed vocal sound). These two features often co-occur in a way that a pressed voice goes along with high tension, but a breathy voice (in the ears of an untrained listener also noise) can go along with low tension, and high or low tension can occur without roughness or breathiness. Hence, the two features fulfill different purposes and can be perceptually differentiated.

The similarity in articulation peculiarities and precision (Function 2) is conceptually not surprising and is grouped with mode of phonation, which may suggest that singing and speaking were evaluated on the basis of articulatory features (already seen in Study 2; see also Merrill & Larrouy-Maestri, 2017). Timbre (squeaky/nasal–dark/dull) and average pitch (low–high) load on the third function, showing that a high pitch in the selected voices is accompanied by a squeaky/nasal sound (e.g., Jackson) and a low pitch with a dark/dull pitch (e.g., Presley). A mix can be found in Houston, which tends toward a squeaky (shrill/bright) sound in higher notes and a darker sound in lower notes, which is reflected in the middle ratings. Expression and pitch changes load on the last function, which reflects that both assess timely variations (poles of varied–uniform

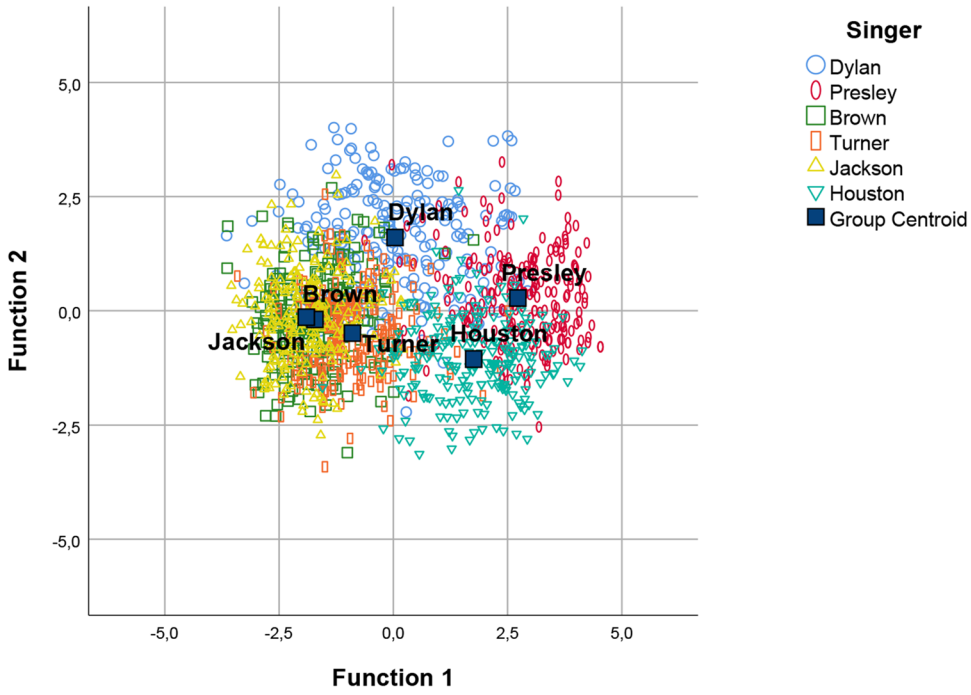

Fig. 5 Scatterplot of the Result of the CDA With Singers and Characteristics. *Note.* On the x-axis, a distinction is made between the features of the first function of the CDA (noise and tension) and those of the second function (articulation and mode of phonation) on the y-axis

and sudden–gliding), which do not contribute to the discrimination of the singers.

The CDA was performed for the German and the English groups separately. Classification results were comparable and support the notion that the features are perceived similarly between the languages (Tables S9 and S10).

Discussion

The present series of studies revealed that untrained listeners can describe singing voices in popular music using meaningful vocal characteristics. Even though free descriptions of voices by untrained listeners were limited to a few descriptions, they were able to assess voices in a similar manner to trained listeners using features from essential categories of vocal description such as sound, pitch, articulation, and overall expression (e.g., Bose, 2001; Wapnick & Ekholm, 1997). The results showed that neither language nor expertise affected the assessment of the singers using the nine items in the German and English questionnaire.

An increase in interrater agreement was seen from the second study to the online survey, showing that the adjustments helped the participants to use the items. The interrater agreement lead to satisfactory results by being comparable to other studies using untrained listeners (Bänziger et al., 2014) and other larger questionnaires comparable to the current study (Merrill & Larrouy-Maestri, 2017). The ICC should increase when the number of categories is reduced, for example, focusing on features related to the sound of voice and including training (as seen in the evaluation of roughness and breathiness; Anders et al., 1988). Nonetheless, some features lead to indecisive ratings in some singers, reflecting the ambiguous usage of that feature by the singer rather than listener unreliability (Kreiman et al., 2007).

The voice profiles captured with the nine features in the current study led to a representation of the six singers in their characteristic properties as shown with a discriminant analysis. Important for the discrimination was the feature roughness which has been shown to be a salient feature of untrained listeners' perceptions (Anders et al., 1988; Bänziger et al., 2014). It should be noted that descriptions by trained listeners will lead to far more detailed descriptions and therefore, probably to a better discrimination. In future research, the classification based on these nine items should also be extended to a larger set of voices.

Feature Descriptions

The chosen procedure generated a selection of voices that originated from the participants' everyday musical life, and essential characteristics to be evaluated crystallized out, which appeared in the free description and thus in the regular vocabulary of the untrained listeners. The resulting set of features used in the online survey entailed features seen in previous studies and theoretical categories of voice description, which will be discussed in detail in the following.

The feature "noise" (rough/scratchy–soft) represents a most important and striking vocal sound, which determines, together with tension, the discrimination of the singers in the current study. Besides its role in clinical research indicating

voice disorder (Hirano, 1981; Mathieson, 2001), it has to be considered an essential vocal feature in the aesthetics of popular vocal performances (and is highly disputed; see Hähnel, 2015). Studies have shown that untrained listeners can only poorly distinguish between roughness and breathiness (Kreiman & Gerratt, 1998), or need training in order to do so (Anders et al., 1988). Adding the term "scratchy" seemed to have given enough information about what the feature described.

Another striking feature in the sound of the singing voice is twang (Hähnel, 2015; Henrich et al., 2008), which is characterized by a specific change in the sound spectrum (Sundberg & Thalén, 2010), and in the present study is represented in the feature-complex timbre (squeaky/nasal–dark/dull). It can be accompanied by a nasal sound (Henrich et al., 2008), which was a common term among untrained listeners, but is often misused and associated with a squeaky sound, making their combination eligible. The term "squeaky" was adopted because of its onomatopoeic effect and because it was used by the untrained listeners.

The term "pressed" was mentioned by the participants and is represented in other singing inventories (Henrich et al., 2008). The opposing characteristic was "pressureless," altogether reflecting "tension," i.e., subglottic air pressure, which varies between different singing styles (Thalén & Sundberg, 2001) and was shown to be an important item for the discrimination of singers.

The feature "average pitch," i.e., the perceived fundamental frequency, reflects a lower or higher voice and is a common feature in tools evaluating the speaking voice (e.g., Bänziger et al., 2014; Mathieson, 2001). In assessments of the singing voice, it has been evaluated in the context of timbre (Henrich et al., 2008) because acoustically a bright or dark timbre are characterized by a lower/higher formant spectrum (Sundberg, 1975) and leads to the impression of a higher or lower voice. In the current study, this feature has led to differences in agreement in the assessed voices, probably because the evaluation depends on the pitch range of a song, and as is to be expected, the melody influences the evaluation of average pitch and is also evaluated in this dependence with music (Colton, 1987).

Due to the overlaps between popular music singing styles and speech, a feature that simply catches the impression of song or speech was included in the final set (mode of phonation; Bose, 2001), which contributed (slightly) to the discrimination of the singers in Study 3. In the beginning, the feature was created to address shouting as a form of singing as it was uttered by the participants in the interview study (and exists as a defining feature in certain musical styles). Because participants in Study 2 did not find this sufficient to describe the voices, the pole was changed to speaking. This features is variable and can be adjusted according to the research questions (speaking, singing, shouting) and it can be investigated how the impression of singing and speaking relates to other features in the questionnaire (Merrill, 2017; Merrill & Larrouy-Maestri, 2017).

Related to mode of phonation is the feature "pitch changes" (sudden–gliding). This feature can be considered necessary for the singing voice profiles because it describes the use of a glissando or portamento in musical terms. In this form, it relates to another singing evaluation where it falls under the category of "melodic articulation" (Henrich et al., 2008). It might play a role in the context of popular music expressions such as crooning or moaning (Hähnel, 2015).

Of all the aspects of articulation that can be evaluated in speech (e.g., Bose, 2001; Laver, 1980) and song (duration of sounds, stress; e.g., Garnier et al., 2007; Henrich et al., 2008), an important one has been intelligibility (e.g., evaluated in the GVPS with "good–bad"; Bänziger et al., 2014). While in a professional singing evaluation, these details may be used to reflect peculiarities of a singer, in the current investigation, the articulation was queried on the level of precision and peculiarities. While precision reflects mainly intelligibility, the feature peculiarities can be used to describe a singer in terms of mannerism, without any specifics, which was important to the participants in Study 2.

As the evaluation of a voice is an evaluation of timely variation, the feature vocal expression ("varied–uniform") enables an evaluation of an overall impression of the vocal performance, which can reflect and possibly explain the ratings of single items. For example, if a voice is evaluated as "varied," it might explain why other single items were rated indecisively.

Limitations

In order to use the current list of vocal features as a standardized tool, psychometric properties will need to be evaluated with a large set of voices, because only a small number was presented in the current study. Further, a 5-point Likert scale is recommended for the ratings, but was not evaluated with the current study, because an even-point scale was used in order to reveal tendencies of the bipolar items. All evaluated voices were presented with music, which might have influenced the ratings of vocal features, e.g., covered up some aspects of the vocal sound. A Capella voices will be interesting to investigate in future research, also because it will enable a comparison of auditory and acoustical features, which was not possible in the current study because of the music in the background.

Conclusion and Outlook

The present series of studies revealed that people having different voice experience are able to describe voices on fundamental categories of vocal expression when presented

with a suitable set of vocal features. The set developed in the present study allows for the creation of comprehensive vocal profiles that reflect a listener's perception of a singing voice, which does not require any previous education or training for its application and is therefore a valuable addition to existing inventories.

The implication of this finding is that untrained listeners are now enabled to communicate about vocal expression in singing, which gives the opportunity for various applications. The emerging vocal profile (or single features from the list) can guide the identification and perceptual evaluation of vocal features during listening, and it can be a useful discussion aid for trained and untrained listeners from different fields (e.g., between students and teachers in music schools, clinical personnel and patients, or voice researchers and study participants). It will help to understand inference mechanisms in vocal communication because it can be used to investigate the connections between vocal characteristics and certain reactions and sensations that occur when singing voices are heard (e.g., emotional expression in song or aesthetic judgements). The features are applicable when acoustical analyses are not possible (e.g., because of background music in the stimulus) or when more subjective perceptions of the individual are of interest.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12144-022-02734-7.

Acknowledgments I would like to thank Taren-Ida Ackermann for help with planning and conducting the interviews; Sandro Wiesmann for help with data collection in the interview and group-testing sessions; Freya Materne and Claudia Lehr for managing the participants; Lutz–Christian Anders and Clara Finke for support with the questionnaire; and Pauline Larrouy–Maestri, Lauren Fink, and Sven Grawunder for critical feedback on an earlier version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data Availability The questionnaire data for these studies are available as supplementary material.

Declarations

Conflict of Interest I state that there is no conflict of interest.

References

Ackermann, T.-I. (2019). *Disliked Music: Merkmale, Gründe und Funktionen abgelehnter Musik*. Kassel University Press. https://doi.org/10.17170/kobra-202007161459.

Ackermann, T.-I., & Merrill, J. (2021). *Rationales and functions of disliked music: An in-depth interview study.* https://doi.org/10.31234/osf.io/5zuwp.

Anders, L. C., Hollien, H., Hurme, P., Sonninen, A., & Wendler, J. (1988). Perception of hoarseness by several classes of listeners. *Folia Phoniatrica et Logopaedica, 40*(2), 91–100. https://doi.org/10.1159/000265889

Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PLoS ONE, 9*(2), e88616. https://doi.org/10.1371/journal.pone.0088616

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636. https://doi.org/10.1037//0022-3514.70.3.614

Bänziger, T., Patel, S., & Scherer, K. R. (2014). The role of perceived voice and speech characteristics in vocal emotion communication. *Journal of Nonverbal Behavior, 38*(1), 31–52. https://doi.org/10.1007/s10919-013-0165-x

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370. https://doi.org/10.1037//1089-2680.5.4.323

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100*(1), 100–130. https://doi.org/10.1016/j.cognition.2005.11.007

Bose, I. (2001). Methoden der Sprechausdrucksbeschreibung am Beispiel kindlicher Spielkommunikation. *Gesprächsforschung - Online-Zeitschrift Zur Verbalen Interaktion, 2*, 262–303 http://www.gespraechsforschung-ozs.de/heft2001/ga-bose.pdf

Cleveland, T. F., Sundberg, J., & Stone, R. (2001). Long-term-average spectrum characteristics of country singers during speaking and singing. *Journal of Voice, 15*(1), 54–60. https://doi.org/10.1016/S0892-1997(01)00006-6

Colton, R. H. (1987). The role of pitch in the discrimination of voice quality. *Journal of Voice, 1*(3), 240–245. https://doi.org/10.1016/S0892-1997(87)80006-1

Ekholm, E., Papagiannis, G. C., & Chagnon, F. P. (1998). Relating objective measurements to expert evaluation of voice quality in Western classical singing: Critical perceptual parameters. *Journal of Voice : Official Journal of the Voice Foundation, 12*(2), 182–196. https://doi.org/10.1016/s0892-1997(98)80038-6

Garnier, M., Henrich, N., Castellengo, M., Sotiropoulos, D., & Dubois, D. (2007). Characterisation of voice quality in western lyrical singing: From teachers' judgements to acoustic descriptions. *Journal of Interdisciplinary Music Studies, 1*(2), 62–91 https://hal.archives-ouvertes.fr/hal-00204134

Greasley, A., Lamont, A., & Sloboda, J. (2013). Exploring musical preferences: An in-depth qualitative study of adults' liking for music in their personal collections. *Qualitative Research in Psychology, 10*(4), 402–427. https://doi.org/10.1080/14780887.2011.647259

Hähnel, T. (2015). Was ist populärer Gesang? Zur Terminologie vokaler Gestaltungsmittel in populärer Musik. In M. Pfleiderer, T. Hähnel, K. Horn, & C. Bielefeldt (Eds.), *Texte zur populären Musik: Vol. 8. Stimme, Kultur, Identität: Vokaler Ausdruck in der populären Musik der USA, 1900–1960* (pp. 53–74) Transcript.

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language, 88*, 70–86. https://doi.org/10.1016/j.jml.2016.01.001

Henrich, N., Bezard, P., Expert, R., Garnier, M., Guerin, C., Pillot-Loiseau, C., Quattrocchi, S., Roubeau, B., & Terk, B. (2008). Towards a common terminology to describe voice quality in western lyrical singing: Contribution of a multidisciplinary research group. *Journal of Interdisciplinary Music Studies, 2*(1&2), 71–93 https://hal.inria.fr/IJLRDA/hal-00297248v1

Hirano, M. (Ed.). (1981). *Disorders of human communication: Vol. 5. Clinical examination of voice*. Springer.

Hirano, M. (1989). Objective evaluation of the human voice: Clinical aspects. *Folia Phoniatrica et Logopaedica, 41*(2–3), 89–144. https://doi.org/10.1159/000265950

Hollien, H. (2000). The concept of ideal voice quality. In R. D. Kent & M. J. Ball (Eds.), *Voice quality measurement* (pp. 13–24). Singular Publ. Group.

Kreiman, J., & Gerratt, B. R. (1998). Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America, 104*(3 Pt 1), 1598–1608.

Kreiman, J., Gerratt, B. R., & Ito, M. (2007). When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America, 122*(4), 2354–2364. https://doi.org/10.1121/1.2770547

Larrouy-Maestri, P. (2018). "I know it when I hear it": On listeners' perception of mistuning. *Music & Science, 1*(1), 205920431878458. https://doi.org/10.1177/2059204318784582

Laver, J. (1980). *THE phonetic description of voice quality. Cambridge studies in linguistics: Vol. 31*. Cambridge Univ. Press.

Mathieson, L. (2001). *Greene and Mathieson's the voice and its disorders* (6th ed.). Whurr Publishers Ltd..

Merrill, J. (2017). Schoenberg's Pierrot Lunaire revisited: Acceptance of vocal expression. *Acta Musicologica, 89*(1), 95–117 https://acta.musicology.org/pdfs/archive/acta2017.pdf

Merrill, J. (2019). Perzeption und Rezeption des vokalen Ausdrucks im Grenzbereich von Singen und Sprechen. *Sprechen. Zeitschrift Für Sprechwissenschaft, Sprechpädagogik, Sprechtherapie, Sprechkunst, 68*, 42–58.

Merrill, J., & Ackermann, T.-I. (2020). "Like static noise in a beautiful landscape": A mixed-methods approach to rationales and features of disliked voices in popular music. *Psychology of Aesthetics, Creativity, and the Arts.* Advance online publication. https://doi.org/10.1037/aca0000376.

Merrill, J., & Larrouy-Maestri, P. (2017). Vocal features of song and speech: Insights from Schoenberg's Pierrot Lunaire. *Frontiers in Psychology, 8*, 1108. https://doi.org/10.3389/fpsyg.2017.01108

Mitchell, H. F. (2014). Perception, evaluation and communication of singing voices. In S. D. Harrison & J. O'Bryan (Eds.), *Teaching singing in the 21st century* (pp. 187–200) Spinger.

Nawka, T., Anders, L. C., & Wendler, J. (1994). Die auditive Beurteilung heiserer Stimmen nach dem RBH-System. *Sprache, Stimme, Gehör, 18*, 130–133.

Oates, J. M., Bain, B., Davis, P., Chapman, J., & Kenny, D. (2006). Development of an auditory-perceptual rating instrument for the operatic singing voice. *Journal of Voice : Official Journal of the Voice Foundation, 20*(1), 71–81. https://doi.org/10.1016/j.jvoice.2005.01.006

Reddy, A. A., & Subramanian, U. (2015). Singers' and nonsingers' perception of vocal vibrato. *Journal of Voice : Official Journal of the Voice Foundation, 29*(5), 603–610. https://doi.org/10.1016/j.jvoice.2014.09.022

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Sadolin, C. (2009). *Complete Vocal Technique*. Shout Publications ApS.

Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice, 9*(3), 235–248. https://doi.org/10.1016/S0892-1997(05)80231-0

Stone, R., Cleveland, T. F., Sundberg, J., & Prokop, J. (2003). Aerodynamic and acoustical measures of speech, operatic, and broadway vocal styles in a professional female singer. *Journal of Voice, 17*(3), 283–297. https://doi.org/10.1067/S0892-1997(03)00074-2

Sundberg, J. (1975). Formant technique in a professional female singer. *Acta Acustica united with Acustica, 32*(2), 89–96.

Sundberg, J., & Thalén, M. (2010). What is "Twang"? *Journal of Voice : Official Journal of the Voice Foundation, 24*(6), 654–660. https://doi.org/10.1016/j.jvoice.2009.03.003

Thalén, M., & Sundberg, J. (2001). Describing different styles of singing: A comparison of a female singer's voice source in "classical", "pop", "jazz" and "blues". *Logopedics, Phoniatrics, Vocology, 26*(2), 82–93. https://doi.org/10.1080/140154301753207458

Wapnick, J., & Ekholm, E. (1997). Expert consensus in solo voice performance evaluation. *Journal of Voice, 11*(4), 429–436. https://doi.org/10.1016/S0892-1997(97)80039-2

Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., & Wilson, J. A. (2004). The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Oto-Rhino-Laryngology, 261*(8), 429–434. https://doi.org/10.1007/s00405-003-0707-7

Zuckerman, M., & Miyake, K. (1993). The attractive voice: What makes it so? *Journal of Nonverbal Behavior, 17*(2), 119–135. https://doi.org/10.1007/BF01001960