No difference in vocal emotion perception between non-professional/amateur singers and instrumentalists

Christine Nussbaum1,2, Jessica Senftleben1, Annett Schirmer3,1, and Stefan R. Schweinberger1,2,4

1Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Germany

2Voice Research Unit, Friedrich Schiller University, Jena, Germany

3Institute of Psychology, University of Innsbruck, Austria

4Swiss Center for Affective Sciences, University of Geneva, Switzerland

Correspondence should be addressed to Christine Nussbaum, Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Leutragraben 1, 07743 Jena, Germany. Tel: +49 (0) 3641 945939, E-Mail: christine.nussbaum@uni-jena.de; or to Stefan R. Schweinberger, Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Am Steiger 3/Haus 1, 07743 Jena, Germany. Tel: +49 (0) 3641 945181, Fax: +49 (0)3641 9 45182, E-Mail: stefan.schweinberger@uni-jena.de. Preprocessed data, analysis scripts and supplemental materials can be found in the associated OSF repository (ToDo)

**Word count:** aim 5000

# Abstract

ToDo

Copied from preregistration for overview:

In this study, we plan to compare vocal emotion perception of two groups of non-professional musicians: singers and instrumentalists. Previous work has shown that musicians outperform non-musicians in vocal emotion recognition. This was recently replicated in a study comparing professional musicians with non-musicians (Study 1: https://osf.io/5tczs/, June 2023 - in revision). In Study 1, we further showed that musicians are particularly sensitive to emotional pitch cues in voices, which seems to be due to pre-disposed natural sensitivity towards pitch and melody cues rather than musical training. The present study is designed as a follow-up on Study 1 and addresses two main questions: First, it aims at a comparison between singers and instrumentalists. Currently, the available literature on singers and instrumentalists with regard to vocal emotional processing is inconsistent and requires further systematic investigation. Second, the present study targets non-professional amateur musicians, in contrast to the professional ones recruited for Study 1. There is accumulating evidence for qualitative differences between professional musicians and musicial amateurs. However, it is unclear how these groups diverge with regard to vocal emotion perception. Therefore, we plan to pool the data collected in the present study and the ones from Study 1 to compare the performance of musical amateurs to the one of professional musicians and non-musicians.

**Keywords:** vocal emotion perception, singers, instrumentalists, amateurs, parameter-specific voice morphing, musicality

# Public significance statement

* ToDo

# Introduction

The human voice is a prime carrier of emotional information and adequate perception of vocal emotions is important for everyday social interaction (Quelle). On average, humans can infer emotion from voices well above chance (Quelle), but this capacity is subject to great individual variability and seems to be linked to differences in **musicality**. It has been shown repeatedly that musicians outperform non-musician in vocal emotion recognition, although the overall effect size can be considered small to moderate (Quellen, for systematic reviews …). Several works shed light on the potential mechanisms underlying this advantage and emphasize the role of **acoustic sensitivity**: Musicians and non-musicians seem to differ in a variety of tasks that target basic auditory perception, including … (Baldè, Kraus& Chandrasekraran, etc). Correia et al (2022) found that the link between musicality and vocal emotion perception was fully mediated via auditory perception skills. Some of these could specifically linked to differences in vocal emotion perception. In a previous study, we found that musicians are particularly tuned to emotional cues expressed by the **fundamental frequency (F0)** contour, sometimes also referred to as voice melody (Nussbaum 2024). We employed parameter-specific voice morphing to create vocal stimuli that expressed the emotional information only through F0, timbre or both. F0 is linked to perceived pitch and timbre is linked to perceived quality of a voice (i.e. whether it sounds harsh or gentle). We found that a professional group of musicians outperformed a group of non-musicians when emotions were expressed by F0 and both cues, but not timbre alone. Additional exploratory analysis revealed two interesting findings: First, we observed a correlation between music perception skills (especially for melodies) and vocal emotion recognition, further corroborating the importance of auditory sensitivity. More importantly, however, we even found this correlation in the group of non-musicians only, who never received any formal music training. Although limited by their correlational nature, these findings suggest that musicians’ advantage in vocal emotion perception may not be the result of music training, but to due to predisposed aptitude to exploit melodic patterns. And indeed, this is supported by a recent randomized-controlled study in school children, which found no evidence for a causal effect of musical training on vocal emotion perception performance (quelle). Thus, there is consensus in the literature that the observed performance difference in musicians and non-musicians is due to variations in acoustic sensitivity rather than the result of formal musical education.

However, musicians are not a homogenous group. On the one hand, there are quantitative differences, regarding very different levels of expertise. On the other hand, there are qualitative differences, as musicians display a great variety of styles, genres, and forms of expression, within the scope of the Wester music system and beyond. Studies that treat musicians as one group disregard this heterogeneity.

The capacity to extract emotional cues from voices

## ToDo

ToDo

## ToDo

ToDo

## Rationale, outline etc.

# Part I: Comparison of non-professional singers and instrumentalists

## Hypotheses

**H1:** We expect NO difference between singers and instrumentalists in overall vocal emotion recognition performance.

**H2:** We expect NO difference between singers and instrumentalists in vocal emotion recognition performance based on timbre and F0 cues only.

## Method

Note that this is a follow-up to the study reported in Nussbaum et al. (2024). Thus, the stimulus material and the design are identical, but we recruited a new sample. The present study was preregistered (<https://doi.org/10.17605/OSF.IO/76PV5>).

### Participants

According to our preregistered plan, we aimed at a sample size of 40 singers (20 male, 20 female) and 40 instrumentalists (20 male, 20 female), because in our previous study, this sample size allowed us to reveal medium-sized group effects (d =0.81 for the Full and d = 0.56 for the F0 morphing condition) when we compared professional musicians and non-musicians.

Data were collected in a pseudonymized format from June 2023 to January 2024. All participants were aged between 18 and 54 years and fluent German speakers. Participants provided informed consent before completing the experiment and received compensation in form of 12.50 € or course credit upon completion. The experiment was in line with the ethical guidelines of the German Society of Psychology (DGPs) and approved by the local ethics committee of the Friedrich Schiller University Jena (Reg.-Nr. FSV 19/045).

In total, we collected data from 94 non-professional musicians (further also referred to as amateurs) that were divided into singers and instrumentalists. Recruitment criteria specified that participants had to be non-professional musicians (i.e., they held no music-related academic degree or worked professionally as a musician). Singers were required to be currently active in a choir or another singing group, but should not play an instrument actively and regularly (i.e., they must not currently be in an orchestra or a band). Instrumentalists, conversely, were required to be currently active in an orchestra or a band, but they should not engage in singing activities actively and regularly (i.e., they must not currently be in a choir or another singing group).

*Singers*

We recorded data from 48 singers, of which three were excluded (N = 2 had > 5 % trials of omission, N = 1 had technical issues during stimulus playback). Thus, data from 45 singers were analyzed (22 female, 22 male, aged 18 to 53 years [M = 27.02, SD = 8.2]). Mean onset age of musical training was 8 years (SD = 3.08, 5 - 20 years). Twenty-two participants had over 10 years of training, ten between 6-9 years, four between 4-5 years and three between 2-3 years. Six participants had less than one year of training (for more details see ToDo).

*Instrumentalists*

Data from 46 instrumentalists were collected, of which three were excluded. One had technical issues during stimulus playback, one was also active in a choir, one held a Master’s degree in music science and was therefore relocated into the group of professional musicians (see Part II). Thus, data from 43 instrumentalists entered analysis (24 female, 18 male, aged 18 to 54 years [M = 28.51, SD = 10.64]). Mean onset of musical training was 7 years (SD = 2.27, 4 - 14 years). Thirty-five participants had studied their instrument for over 10 years, one between 6-9 years and three between 1-2 years. Four participants had less than one year of training (for more details see ToDo).

### Stimulus material

As stimulus material, we used parameter-specific voice morphs that express the emotional information either through the fundamental frequency contour only (F0), through timbre only (Tbr) or through a combination of both (Full).

For voice morphing, we selected original audio recordings from a database of vocal actor portrayals, comprised of pseudowords (/molen/, /loman/, /belam/) uttered by eight speakers (four male, four female) with expressions of happiness, pleasure, fear, and sadness. To synthesize the parameter-specific emotional voice morphs, we created morphing trajectories between each emotion and an emotional average of the same speaker and pseudoword, using the Tandem-STRAIGHT software (Kawahara et al., 2013; Kawahara et al., 2008). The averages had been created previously by blending all emotions together and were thus assumed to be uninformative and unbiased with respect to the four emotions of interest. After substantial preprocessing (e.g. manual mapping of time- and frequency anchors in each stimulus), Tandem-STRAIGHT enables voice morphing via weighted interpolation of five independent parameters: (1) F0-contour, (2) timing, (3) spectrum-level, (4) aperiodicity, and (5) spectral frequency; the latter three are summarized as timbre.

We created three types of morphed stimuli (see **Figure 1**). **Full-Morphs** were stimuli with all parameters taken from the emotional version (corresponding to 100% from the emotion and 0% from average), except for the timing parameter, which was taken from the average (corresponding to 0% emotion and 100% average). **F0-Morphs** were stimuli with the F0-contour taken from the emotion, but timbre and timing taken from the average. **Timbre-Morphs** were stimuli with all timbre parameters taken from the emotion, but F0 and timing from the average. Note that the timing was kept constant across all conditions to allow a pure comparison of F0 vs. timbre. Furthermore, we kept all average stimuli as a further ambiguous reference category. In total, this resulted in 8 (speakers) x 3 (pseudowords) x 4 (emotions) x 3 (morphing conditions) + 24 average (8 speakers x 3 pseudowords) = 312 stimuli (duration M = 780 ms, range 620 to 967 ms, SD = 98 ms). Using PRAAT (Boersma, 2018), we normalized all stimuli to a root-mean-square of 70 dB SPL.

For a more detailed description of the stimulus creation, see Nussbaum et al. (2024). For general information about voice morphing, refer to Kawahara and Skuk (2019).

**Figure 1**

*Morphing matrix for stimuli with averaged voices as reference*



*Note. Figure reprinted from Nussbaum et al. (2024)*

### Design

Data were collected online via PsyToolkit (Stoet, 2010, 2017), but after completion of the study all participants met with the experimenter for a short personal debriefing. This was done to increase commitment and conscientiousness for the experiment.

Participants were required to ensure a quiet environment for the duration of the study and use a computer with a physical keyboard and headphones. As browser, we recommended Google Chrome, and excluded Safari for technical reasons. Prior to the listening tasks, participants could adjust their sound settings to a comfortable sound pressure level

First, participants entered demographic information, including age, sex, native language, profession, and potential hearing impairments such as tinnitus. They then completed an emotion classification experiment, a test on music perception (the Profile of Music Perception Skills) and several questionnaires on musicality, personality and socioeconomic background. Mean duration of the whole online study was about 75 minutes.

Emotion classification experiment

In the experiment, participants classified vocal emotions as happiness, pleasure, fear, or sadness. Each trial started with a green fixation cross presented for 500 ms. Then the sound was played while a loudspeaker symbol was shown on the screen. Subsequently, a response screen showed the emotion labels and participants could enter their response within a 5000 ms time window starting from voice offset. Responses were entered with the left and right index and middle fingers, with random mapping of response keys to emotion categories for each participant, out of four possible key mappings (see Tables on OSF ToDo). In case of a response omission, the final trial slide (500 ms) prompted participants to respond faster; otherwise, the screen turned back. Then the next trial started.

Preceding the experiment, participants completed eight practice trials with stimuli not used during the actual task. Subsequently, all 312 experimental stimuli were presented once in randomized order across six blocks of 52 trials each. Between blocks, participants could take self-paced breaks. The total duration of the experiment was about 25 minutes.

Profile of Music Perception Skills (PROMS)

To measure music perception skills, we used a modular version of the Profile of Music Perception Skills (Law & Zentner, 2012; Zentner & Strauss, 2017), comprised of the four subtests „Melody“, „Pitch“, „Timbre”, and „Rhythm“. Participants completed 18 items per subtest, always preceded by one practice trial. Each trial, participants heard a reference stimulus twice followed by a target stimulus. Then, they indicated whether reference and target were the same or different via a 5-point Likert scale with the labels “definitely same”, “maybe same”, “don’t know”, “maybe different”, and “definitely different”. Duration of the PROMS was about 20 minutes.

Questionnaires

After the PROMS, participants completed several questionnaires: the German Version of the Autism Quotient Questionnaire, AQ, (Baron-Cohen et al., 2001; Freitag et al., 2007), a 30-item Personality Inventory measuring the Big-Five domains (Rammstedt et al., 2018), the Goldsmiths Musical Sophistication Index, Gold-MSI, (Müllensiefen et al., 2014) to assess the participants’ degree of self-reported musical skills, additional questions concerning music experience and musical engagement, their socioeconomic background, and the 20-item version of the Positive-Affect-Negative-Affect-Scale, PANAS (Breyer & Bluemke, 2016; Watson et al., 1988).

### Data analysis

In line with our preregistered plan, we collapsed data across speakers and pseudowords for analysis. Further, data on emotional averaged were excluded because they were not relevant for our hypotheses. Response omissions (~1%) were treated as errors and participants with more than 5% of such omissions excluded from data analysis. Analyses of Variance (ANOVAs) and correlational analyses were performed using R Version 4.3.2 (R Core Team, 2020). Post-hoc tests were Benjamini-Hochberg corrected where appropriate (Benjamini & Hochberg, 1995).

We complemented these classical frequentist analyses with a Bayesian approach, which – in contrast to null hypothesis significance testing - allows a quantification of evidence for null findings (Rosenfeld & Olson, 2021). These analyses were conducted in JASP Version 0.19.3 (Quelle) using default priors. We report the Bayes factor (BF10) as an indicator for the likelihood of the null and alternative hypothesis given the observed data. BF10 > 1 indicate larger evidence for the alternative hypothesis, BF10 < 1 larger evidence for the null hypothesis. For example, a BF10 = 3 means that the alternative hypothesis is three times more likely than the null hypothesis, whereas the reciprocal BF = .33 means that the null hypothesis is three times more likely than the alternative. Following the guidelines by Jarosz and Wiley (2014) we consider values of BF10 = 1-3 (.1-.33) as anecdotal, BF10 = 3-10 (.33-.10) as moderate, BF10 = 10-30 (.10-.03) as strong, BF10 = 30-100 (.03-.01) as very strong and BF10 > 100 (< .01) as decisive evidence for the alternative hypothesis and the reciprocal values in parenthesis as respective evidence for the null hypothesis.

In alignment with the approach by Nussbaum et al. (2024), we recoded responses in the PROMS from 0 to 1 in 0.25 steps starting with the “definitely” correct option down two the “definitely” incorrect option (thus, “don’t know” was always coded with 0.5) and subtracted 0.5 from the final measure. Thus, a positive score indicates that participants were more correct/confident, whereas a negative score indicates more incorrect/uncertain ratings. For statistical analyses, we used the averaged performance across trials for each subtest.

## Transparency and openness

We specified how we determined our sample size, all data exclusions, all manipulations, and all measures in the associated preregistration (<https://doi.org/10.17605/OSF.IO/76PV5>). Preprocessed data, analysis scripts and supplemental materials can be found in the associated OSF repository (ToDo). For stimulus examples, refer by Nussbaum et al. (2024).

## Results

### Demography, musicality, and personality of participants

Singers and instrumentalists did not differ significantly in the socioeconomic status assessed via educational level, X2(2, N = 88) = 1.06, p = .588; highest academic degree, X2(7, N = 88) = 9.06, p = .249, and household income, X2(4, N = 88) = 5.23, p = .264 (for more details see Table ToDo). Further, the groups did not differ in age or positive and negative affect and were comparable regarding Big Five personality traits and autistic traits. In the Gold-MSI, singers and instrumentalists scored comparatively on the general musicality score, but there were differences on two subfactors: instrumentalists scored higher on the subfactor Formal Education, while singers scored higher on Singing. In the PROMS, both groups performed comparably in all four subtests. Participant characteristics assessed via self-report and music performance in the PROMS are summarized in **Table 1**.

**Table 1**

*Characteristics of participants - Demography, personality, and musicality*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Singers** | **Instrumentalists** |  |  |  |  |  |
|  | M (SD) | M (SD) | t | dfa | p | Cohens d |  |
| Age | 27.02 (8.2) | 28.51 (10.6) | -0.73 | 78.93 | .465 | -0.17 [-0.61, 0.28] |  |
|  |  |  |  |  |  |  |  |
| *PANAS* |  |  |  |  |  |  |  |
| positive Affect | 3.00 (0.68) | 3.11 (0.57) | -0.77 | 84.78 | .446 | -0.17 [-0.59, 0.26] |  |
| negative Affect | 1.53 (0.47) | 1.40 (0.35) | 1.49 | 80.61 | .141 | 0.33 [-0.11, 0.77] |  |
|  |  |  |  |  |  |  |  |
| *Big Five* |  |  |  |  |  |  |  |
| Openness | 4.04 (0.55) | 3.99 (0.51) | 0.46 | 85.96 | .647 | 0.10 [-0.32, 0.52] |  |
| Conscientiousness | 3.47 (0.69) | 3.76 (0.70) | -1.91 | 85.62 | .060 | -0.41 [-0.84, 0.02] |  |
| Extraversion | 3.21 (0.70) | 3.00 (0.73) | 1.44 | 85.3 | .155 | 0.31 [-0.12, 0.74] |  |
| Agreeableness | 3.81 (0.57) | 4.01 (0.60) | -1.61 | 85.29 | .112 | -0.35 [-0.77, 0.08] |  |
| Neuroticism | 2.74 (0.77) | 2.61 (0.78) | 0.80 | 85.75 | .426 | 0.17 [-0.25, 0.60] |  |
|  |  |  |  |  |  |  |  |
| *AQ* |  |  |  |  |  |  |  |
| Total | 18.2 (6.15) | 19.28 (8.55) | -0.68 | 76.04 | .500 | -0.16 [-0.60, 0.30] |  |
| Attention to Detail | 5.4 (2.33) | 5.63 (2.53) | -0.44 | 84.64 | .662 | -0.10 [-0.52, 0.33] |  |
| Social | 12.8 (5.37) | 13.65 (7.53) | -0.61 | 75.65 | .545 | -0.14 [-0.59, 0.31] |  |
| Social Skills | 2.4 (1.94) | 3.09 (2.95) | -1.30 | 72.01 | .200 | -0.31 [-0.77, 0.16] |  |
| Communication | 2.53 (1.94) | 2.44 (2.3) | 0.20 | 82.02 | .842 | 0.04 [-0.39, 0.48] |  |
| Imagination | 2.51 (1.75) | 2.81 (1.88) | -0.78 | 84.86 | .437 | -0.17 [-0.60, 0.26] |  |
| Attention Switching | 5.36 (1.91) | 5.30 (2.23) | 0.12 | 82.69 | .904 | 0.03 [-0.40, 0.46] |  |
|  |  |  |  |  |  |  |  |
| *Gold-MSI* |  |  |  |  |  |  |  |
| General ME | 4.78 (0.85) | 4.75 (0.80) | 0.19 | 85.99 | .866 | 0.04 [-0.39, 0.46] |  |
| Active Engagement | 3.83 (0.82) | 4.21 (1.13) | -1.79 | 76.79 | .078 | -0.41 [-0.86, 0.05] |  |
| Formal Education | **4.39 (1.14)** | **4.95 (0.62)** | -2.85 | 68.31 | .006 | -0.69 [-1.18, -0.20] | \*\* |
| Emotion | 5.50 (0.81) | 5.60 (0.76) | -0.60 | 85.99 | .549 | -0.13 [-0.55, 0.29] |  |
| Singing | **4.98 (0.97)** | **4.19 (1.27)** | 3.25 | 78.56 | .002 | 0.73 [0.27, 1.19] | \*\* |
| Perception | 5.73 (0.82) | 5.77 (1.03) | -0.22 | 80.11 | .825 | -0.05 [-0.49, 0.39] |  |
|  |  |  |  |  |  |  |  |
| *PROMS* |  |  |  |  |  |  |  |
| Pitch | 0.23 (0.08) | 0.24 (0.06) | -0.30 | 82.92 | .766 | -0.07 [-0.50, 0.37] |  |
| Melody | 0.17 (0.10) | 0.14 (0.10) | 1.29 | 85.08 | .199 | 0.28 [-0.15, 0.71] |  |
| Timbre | 0.29 (0.08) | 0.3 (0.09) | -0.59 | 86.00 | .556 | -0.13 [-0.55, 0.30] |  |
| Rhythm | 0.31 (0.09) | 0.30 (0.08) | -0.56 | 85.13 | .577 | -0.12 [-0.55, 0.30] |  |

*Note. Descriptive values show mean ratings for the PANAS (Breyer & Bluemke, 2016), the Big-Five Domains (Rammstedt et al., 2018), and the Gold-MSI (Müllensiefen et al., 2014). AQ score were calculated based on Hoekstra et al. (2008) and Baron-Cohen et al. (2001).*

*a Note that original degrees of freedom were 86 but were corrected due to unequal variance.*

### Emotion classification performance

The mean proportion of correct responses was submitted to an ANOVA with Emotion (Happiness, Pleasure, Fear, and Sadness) and Morph Type (Full, F0, and Timbre) as repeated measures factors and Group (singers and instrumentalists) as a between subject factor (see **Table 1**).

The results revealed main effects of **Emotion** and **Morph Type**, which were qualified by an interaction. Crucially, however, we found no main effects or interactions involving **Group** (see **Figure 2**). Planned Bayesian analysis revealed moderate evidence for the null effect of group for overall performance (p = .542, BF10 = 0.265), as well as for Full (p = .392, BF10 = 0.31), F0 (p = .935, BF10 = 0.226), and Timbre morphs (p = .555, BF10 = 0.262) separately. Thus, we found evidence consistent with our hypotheses H1 and H2.

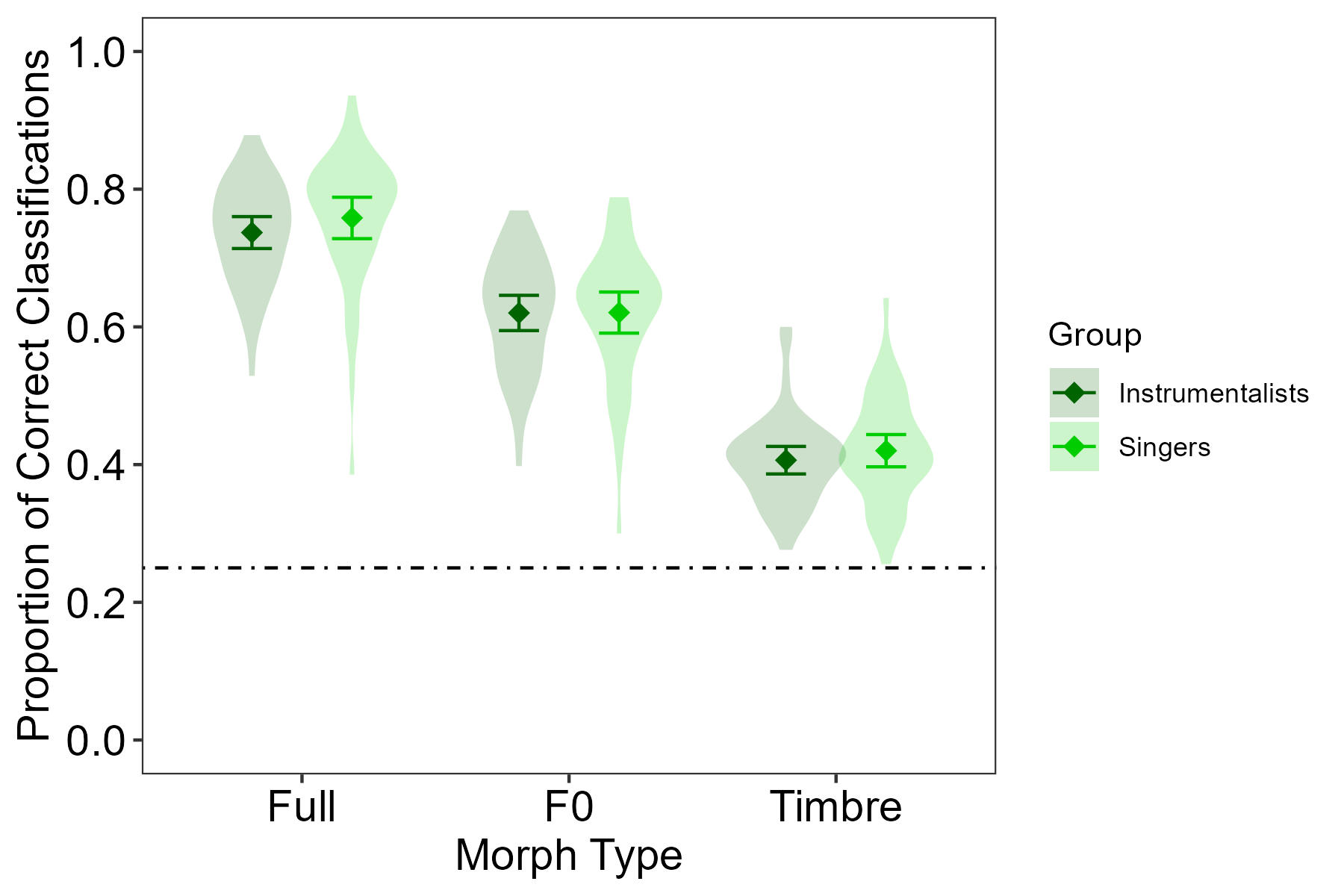
**Table 1**

*Results of the 4 × 3 × 2 mixed-effects ANOVA on the mean proportion of correct responses*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **df1** | **df2** | **F** | **p** | **Ωp2 [95%-CI]** | ***ε*HF** |
| Group | 1 | 86 | 0.38 | .542 | .00 [.00 .01] |  |
| Emotion | 3 | 258 | 72.43 | <.001 | .45 [.36 .53] |  |
| Morph Type | 2 | 172 | 768.93 | <.001 | .90 [.87 .93] | .741 |
| Group x Emotion | 3 | 258 | 2.14 | .095 | .01 [.00 .04] |  |
| Group x Morph Type | 2 | 172 | 0.36 | .635 | .00 [.00 .01] |  |
| Emotion x Morph Type | 6 | 516 | 22.78 | <.001 | .20 [.14 .25] | 827 |
| Group x Emotion x Morph Type | 6 | 516 | 1.33 | .249 | .00 [.00 .01] |  |

**Figure 3**

Mean proportion of correct responses per Morph Type separately for singers and instrumentalists



Note. Whiskers represent 95% confidence intervals. Violin plots represent variation of individual participants. The dotted line represents guessing rate at .25.

Follow-up analysis of the Morph Type effect revealed that performance was best in the Full condition (*M* = 0.75 ± 0.02 SEM), followed by the F0 (*M* = 0.62 ± 0.02) and then the Timbre condition (*M* = 0.41 ± 0.01); Full vs. F0: |*t*(87)| = 23.28, *p* < .001, *d* = 2.50 [2.07, 2.92], F0 vs Timbre: |*t*(87)| = 21.44, *p* < .001, *d* = 2.30 [1.90, 2.70], Full vs Timbre: |*t*(87)| = 33.94, *p* < .001, *d* = 3.64 [3.06, 4.21]). This Morph Type main effect was also found for all emotions separately (all *F*s(2, 174) > 102.44, *p* < .001), although it differed slightly between emotions, as suggested by the interaction (see **Figure 3,** for all post-hoc tests, refer to OSF).

To address our specific interest in the relative importance of F0 and Timbre for the different emotions, we calculated the performance differenceF0-Tbr for each emotion separately. Performance difference was largest for Happiness (M = 0.34 ± 0.02 SEM), followed by Fear (M = 0.21 ± 0.02), Sadness (M = 0.18 ± 0.02), and Pleasure (M = 0.10 ± 0.02; all pairwise comparisons |*t*s(77)| ≥ 2.57, pscorrected ≤ .012, ds ≥ 0.28 [0.06 0.49], except for Fear vs. Sadness (|t(87)| = 1.13, pcorrected = .261).

These effects of Morph Type and Emotion therefore present a full replication of the patterns reported in Nussbaum et al. (2024).

**Figure 3**

Mean proportion of correct responses per Emotion and Morph Type



Note. Whiskers represent 95%-confidence intervals. Grey dots represent individual participants’ data. The dotted line represents guessing rate at .25.

ToDo: additional analyses like confusion data or classification of averages on OSF

# Part II: Comparison of professionals, amateurs and non-musicians

ToDo

## Hypotheses

**H3:** Amateur musicians outperform non-musicians in vocal emotion perception, in the Full and in the F0 condition.

**H4:** Amateurs perform equal or better to professional musicians in the Full and the F0 condition.

# Discussion

ToDo

## ToDo

ToDo

## ToDo

ToDo

## Constraints on generality and future directions

ToDo

# Summary and Conclusion

ToDo

# Acknowledgements

This study was conducted by J.D. in partial fulfilment of the requirements for a master’s thesis. The original voice recordings that served as a basis for creating our stimulus material were provided by Sascha Frühholz. We thank Hannah Strauß for support with the PROMS. We are grateful to all participants of the study.

# Conflicts of Interests and Funding

The authors declare no conflicts of interests.

# Credit Author Statement

Christine Nussbaum – Conceptualization, Methodology, Software, Visualization, Formal analysis, Writing - Original Draft, Supervision

Jessica Dethloff - Data collection, Formal analysis, Visualization, Writing - Original Draft

Annett Schirmer – Methodology, Writing - Review & Editing, Supervision

Stefan R. Schweinberger – Conceptualization, Writing - Review & Editing, Supervision

# Supplementary material

Supplemental figures and tables, analysis scripts, and raw data can be found on the associated OSF repository (ToDo).

References

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J.‑C., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5–17.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Breyer, B., & Bluemke, M. (2016). *Deutsche Version der Positive and Negative Affect Schedule PANAS (GESIS Panel).* https://doi.org/10.6102/zis242

Freitag, C. M., Retz-Junginger, P., Retz, W., Seitz, C., Palmason, H., Meyer, J., Rösler, M., & Gontard, A. von (2007). Evaluation der deutschen Version des Autismus-Spektrum-Quotienten (AQ) - die Kurzversion AQ-k. *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, *36*(4), 280–289. https://doi.org/10.1026/1616-3443.36.4.280

Hoekstra, R. A., Bartels, M., Cath, D. C., & Boomsma, D. I. (2008). Factor structure, reliability and criterion validity of the Autism-Spectrum Quotient (AQ): a study in Dutch population and patient groups. *J Autism Dev Disord*, *38*(8), 1555–1566. https://doi.org/10.1007/s10803-008-0538-x

Jarosz, A. F., & Wiley, J. (2014). What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving*, *7*(1). https://doi.org/10.7771/1932-6246.1167

Kawahara, H., Morise, M., & Skuk, V. G. (2013). Temporally variable multi-aspect N-way morphing based on interference-free speech representations. *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PLoS One*, *7*(12), e52508. https://doi.org/10.1371/journal.pone.0052508

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PLoS One*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0101091

Nussbaum, C., Schirmer, A., & Schweinberger, S. R. (2024). Musicality - Tuned to the melody of vocal emotions. *Br J Psychol*, *115*(2), 206–225. https://doi.org/10.1111/bjop.12684

Rammstedt, B., Danner, D., Soto, C. J., & John, O. P. (2018). Validation of the short and extra-short forms of the Big Five Inventory-2 (BFI-2) and their German adaptations. *European Journal of Psychological Assessment.* Advance online publication. https://doi.org/10.1027/1015-5759/a000481

Rosenfeld, J. P., & Olson, J. M. (2021). Bayesian Data Analysis: A Fresh Approach to Power Issues and Null Hypothesis Interpretation. *Applied Psychophysiology and Biofeedback*, *46*(2), 135–140. https://doi.org/10.1007/s10484-020-09502-y

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: development and validation of the Short‐PROMS and the Mini‐PROMS. *Annals of the New York Academy of Sciences*, *1400*(1), 33–45. https://doi.org/10.1111/nyas.13410