Vocal Emotion Recognition: A Comparison of Singers and Instrumentalists, Amateurs and Professionals

STAR Methods

## **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**

### **Participants for Part I and II**

(reported in the manuscript in section “Part I - Participants”)

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 the form of 12.50 € or course credit upon completion. The experiment was in line with the ethical guidelines by the American Psychological Association (APA) 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 amateur musicians 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 instrumentalists 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, 1 diverse, 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). Mean duration of musical training was 10 years (SD = 6.59, 0 – 25 years). Five participants reported that they never had any formal musical training. Two participants reported that they had occasional tinnitus, but without any subjective impairments in daily life.

#### 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 regrouped with the professional musicians (see Part III). Thus, data from 43 instrumentalists entered analysis (24 female, 18 male, 1 diverse, aged 18 to 54 years [M = 28.51, SD = 10.64]). Mean onset age of musical training was 7 years (SD = 2.27, 4 - 14 years). Mean duration of musical training was 14 years (SD = 10.00, 0 – 44 years). Four participants reported that they never had any formal musical training, thus were autodidacts.

For more details, please refer to the supplemental sample information in the document “Supplemental\_Tables\_and\_Figures.pdf” in the associated OSF repository (<https://osf.io/ascqx/>): S1 for musical background, S2 for socioeconomic background.

### **Participants for Part III**

(reported in the manuscript in section “Part III - Method”)

For this analysis, we collapsed all participants from Part I into the group of amateur musicians and compared it to the groups of professional musicians and non-musicians reported in Nussbaum et al. (2024). Note that we added one participant to the professional group, because he held a master’s degree in music (see Part I), so numbers slightly diverge from the original publication. All professional musicians reported to have a music-related academic degree or a non-academic music qualification. Non-musicians played no instrument or engaged in any other musical activities. For a more detailed description, please refer to Nussbaum et al. (2024).

In total, we analyzed data from 40 professional musicians (20 male, 20 female, aged 20 to 42 years [M = 29.6; SD = 5.58]), 38 non-musicians (18 male, 20 female, aged 19 to 48 years [M = 30.5; SD = 6.54]) and 88 amateurs (40 male, 46 female, 2 diverse, aged 18 to 54 years [M = 27.8; SD = 9.44]).

## **METHOD DETAILS**

### **Stimulus material**

(reported in the manuscript in section “Part I – Stimulus material”)

As stimulus material, we used parameter-specific voice morphs that express 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. We specifically opted for two positive and two negative emotions of different intensities, to balance both valence and arousal. A prior validation study with 20 raters confirmed that the two positive and two negative emotions had different degrees of emotional intensity (happiness > pleasure, t(19)=9.57, p < 0.001 and fear > sadness, t(19)=6.58, p < 0.001).

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. 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 were 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 in 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) and Kawahara and Skuk (2018).

For a summary of acoustic characteristics, see Tables S3 and S4 in the document “Supplemental\_Tables\_and\_Figures.pdf” in the associated OSF repository (<https://osf.io/ascqx/>)

### **Design**

(reported in the manuscript in section “Part I – 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 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 S5 and S6 in the document “Supplemental\_Tables\_and\_Figures.pdf” in the associated OSF repository). In case of a response omission, the final trial slide (500 ms) prompted participants to respond faster; otherwise, the screen turned black. Then the next trial started.

The 312 stimuli were presented in randomized order in six blocks of 52 trials each, with self-paced breaks in between. Beforehand, participants completed eight practice trials with different stimuli. The experiment was about 25 minutes long. Unfortunately, due to a software error, randomization was sampled with replacements, so that some stimuli were drawn repeatedly and others were omitted. This was in contrast to our previous study, where randomization was sampled without replacement so that each stimulus was drawn exactly once.

#### 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“, which we considered most informative for the present research question. 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”. The 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).

## **QUANTIFICATION AND STATISTICAL ANALYSIS**

(reported in the manuscript in section “Part I – Data analysis”)

In line with our preregistered plan, we collapsed data across speakers and pseudowords for analysis. Further, data on emotional averages 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 4.5.0 (R Core Team, 2025). Post-hoc tests were Benjamini-Hochberg corrected where appropriate (Benjamini & Hochberg, 1995). A p-value < 0.05 was considered significant.

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 (JASP Team, 2025) using default priors, which have been considered appropriate for testing null hypotheses based on similar sample sizes (Ly et al., 2016; Neves et al., 2025). Further, we ensured that our Bayesian inference did not depend critically on the choice of priors by running robustness checks (see data analysis files “Analysis\_JASP\_Bayesian.zip” on OSF). 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 indicate 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 BF10 = .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 parentheses as respective evidence for the null hypothesis.

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

## **ADDITIONAL RESOURCES**

(reported in the manuscript in section “Part I – 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 (https://osf.io/ascqx/).

## **KEY RESOURCES TABLE**

**Key resources table**

|  |  |  |
| --- | --- | --- |
| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
| Deposited data | | |
| a zip-Folder with Stimulus examples: “stimulus\_examples.zip” | Associated OSF repository provided by the authors | https://osf.io/ascqx/ |
| a zip-Folder containing the frequentist analysis + preprocessed data in R: “Analysis\_R\_Frequentist.zip” | Associated OSF repository provided by the authors | https://osf.io/ascqx/ |
| a zip-Folder containing the Bayesian analysis in JASP: “Analysis\_JASP\_Bayesian.zip” | Associated OSF repository provided by the authors | https://osf.io/ascqx/ |
| a PDF with supplemental Figures and Tables: “Supplemental\_Tables\_and\_Figures.pdf” | Associated OSF repository provided by the authors | https://osf.io/ascqx/ |
| Software and algorithms | | |
| R Version 4.5.0 (used for statistical analysis) | R Core Team (2025) | https://cran.r-project.org/ |
| JASP Version 0.19.3 (used for statistical analysis) | JASP Team (2025) | https://jasp-stats.org/ |

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

Boersma, P. (2018). Praat: Doing phonetics by computer [Computer program]: Version 6.0.46, retrieved January 2020 from http://www.praat.org/. *Http://www. Praat. Org*.

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

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

JASP Team. (2025). *JASP (Version 0.19.3)[Computer software]*. https://jasp-stats.org/

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*.

Kawahara, H., & Skuk, V. G. (2018). Voice Morphing. In S. Frühholz, P. Belin, S. Frühholz, P. Belin, & K. R. Scherer (Eds.), *The Oxford Handbook of Voice Perception* (pp. 684–706). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.31

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

Ly, A., Verhagen, J., & Wagenmakers, E.‑J. (2016). Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. https://doi.org/10.1016/j.jmp.2015.06.004

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

Neves, L., Martins, M., Correia, A. I., Castro, S. L., Schellenberg, E. G., & Lima, C. F. (2025). Does music training improve emotion recognition and cognitive abilities? Longitudinal and correlational evidence from children. *Cognition*, *259*, 106102. https://doi.org/10.1016/j.cognition.2025.106102

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

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

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