

Decoding emotions in expressive music performances: A multi-lab replication and extension study

Jessica Akkermans, Renee Schapiro, Daniel Müllensiefen, Kelly Jakubowski, Daniel Shanahan, David Baker, Veronika Busch, Kai Lothwesen, Paul Elvers, Timo Fischinger, Kathrin Schlemmer & Klaus Frieler


To cite this article: Jessica Akkermans, Renee Schapiro, Daniel Müllensiefen, Kelly Jakubowski, Daniel Shanahan, David Baker, Veronika Busch, Kai Lothwesen, Paul Elvers, Timo Fischinger, Kathrin Schlemmer & Klaus Frieler (2018): Decoding emotions in expressive music performances: A multi-lab replication and extension study, *Cognition and Emotion*, DOI: [10.1080/02699931.2018.1541312](https://doi.org/10.1080/02699931.2018.1541312)

To link to this article: <https://doi.org/10.1080/02699931.2018.1541312>



Published online: 08 Nov 2018.



Submit your article to this journal 



Article views: 51



View Crossmark data 



Decoding emotions in expressive music performances: A multi-lab replication and extension study

Jessica Akkermans^a, Renee Schapiro^a, Daniel Müllensiefen^a, Kelly Jakubowski^b, Daniel Shanahan^c, David Baker^c, Veronika Busch^d, Kai Lothwesen^d, Paul Elvers^e, Timo Fischinger^e, Kathrin Schlemmer^f and Klaus Frieler^g

^aDepartment of Psychology, Goldsmiths, University of London, London, UK; ^bDepartment of Music, Durham University, Durham, UK; ^cCollege of Humanities and Social Sciences, Louisiana State University, Baton Rouge, LA, USA; ^dDepartment of Musicology and Music Education, University of Bremen, Bremen, Germany; ^eMusic Department, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany; ^fMusic Department, Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany; ^gInstitute for Musicology, University of Music “Franz Liszt” Weimar, Hamburg, Germany

ABSTRACT

With over 560 citations reported on Google Scholar by April 2018, a publication by Juslin and Gabrielsson (1996) presented evidence supporting performers' abilities to communicate, with high accuracy, their intended emotional expressions in music to listeners. Though there have been related studies published on this topic, there has yet to be a direct replication of this paper. A replication is warranted given the paper's influence in the field and the implications of its results. The present experiment joins the recent replication effort by producing a five-lab replication using the original methodology. Expressive performances of seven emotions (e.g. happy, sad, angry, etc.) by professional musicians were recorded using the same three melodies from the original study. Participants ($N = 319$) were presented with recordings and rated how well each emotion matched the emotional quality using a 0–10 scale. The same instruments from the original study (i.e. violin, voice, and flute) were used, with the addition of piano. In an effort to increase the accessibility of the experiment and allow for a more ecologically-valid environment, the recordings were presented using an internet-based survey platform. As an extension to the original study, this experiment investigated how musicality, emotional intelligence, and emotional contagion might explain individual differences in the decoding process. Results found overall high decoding accuracy (57%) when using emotion ratings aggregated for the sample of participants, similar to the method of analysis from the original study. However, when decoding accuracy was scored for each participant individually the average accuracy was much lower (31%). Unlike in the original study, the voice was found to be the most expressive instrument. Generalised Linear Mixed Effects Regression modelling revealed that musical training and emotional engagement with music positively influences emotion decoding accuracy.

ARTICLE HISTORY

Received 2 May 2018
Revised 23 September 2018
Accepted 7 October 2018

KEYWORDS

Emotion decoding; emotion study; musical training; replication; expressive performance

Emotion and music

Plenty of evidence from research literature over the past decades suggests that music can be used to communicate (e.g. Juslin, 1997) and induce emotions (e.g. Gabrielsson & Juslin, 2003; Juslin & Sloboda, 2001). Music can cause emotion-related physiological reactions such as shivers and goosebumps (Jäncke, 2008;

Ward, 2006), trigger behavioural emotional reactions (Jäncke, 2008; Ward, 2006), can be used to regulate mood (Baumgartner, Lutz, Schmidt, & Jäncke, 2006) and reduce stress (Juslin & Laukka, 2004; Juslin & Västfjäll, 2008). However, it is not very clear why certain pieces of music are highly emotional for one individual and have no effect on others, pointing to the

importance of individual differences in musical emotion perception and induction (Juslin & Laukka, 2004; Thompson & Robitaille, 1992). Due to the role of individual differences, it can be difficult to predict a potential reaction to a piece of music (Daly et al., 2015; Yang, Lin, Su, & Chen, 2008). Individual differences may arise as individuals can give different meanings to musical features such as tempo, pitch, and timbre in different contexts, possibly due to individual listening histories and prior associations. For example, high tempo is typically related to joy and happiness but also to anger and rage (Juslin, 1997). Despite individual differences, the associations between basic musical features and perceived emotions are not totally random but show discernable patterns (Gabrielsson & Juslin, 2003).

Previous studies have investigated whether composers and performers can share emotional “codes” with listeners (i.e. shared use of musical cues for communicating and understanding) for emotional expression in music and found high music emotion decoding accuracy (i.e. listeners accurately perceiving the intended emotion; for review, see Eerola & Vuoskoski, 2011; Juslin & Laukka, 2003). The present study sought to replicate indicators for the accuracy of emotion communication in music from the performer to the listener while also considering individual differences that might influence the accuracy of emotion decoding for perceived emotions in music.

Focusing on the communication of emotion in musical performance, Juslin (1997; Juslin & Sloboda, 2001) adopted three assumptions from a functionalist perspective. The assumptions are (1) that emotion decoding is done using basic emotions (De Gelder & Vroomen, 1996; Juslin, 2013; Juslin & Sloboda, 2001), (2) that these basic emotions (such as happiness, anger, and sadness) are easier to communicate than more ambiguous emotions such as solemnity and tenderness (Gabrielsson & Juslin, 1996; Gabrielsson & Lindström, 2010), and (3) that the communication of emotions is driven by social interactions such as the interaction between mother and infant (Juslin, 1997). Two factors influencing emotion decoding in music have been proposed from the functionalist perspective (Juslin, 1997; Juslin & Sloboda, 2001). The first factor considers the innate brain mechanisms for vocal expression of emotion (Juslin & Sloboda, 2001), implying that there is an intimate relationship between music and the human speaking voice, and that there may be parallels between communicating emotions using the voice in speech and

communicating emotions through music (Escoffier, Zhong, Schirmer, & Qui, 2013). The second factor considers social learning and memories. This factor arises in early development with the interaction of parent and infant. Parents often talk to their infants in a different way than they would to other adults, namely by increasing pitch and contour to allow the infant to learn differences in intonation and be able to decode emotions (Juslin, 1997; Trainor, Austin, & Desjardins, 2000). Cultural influences and variances in exposure to social learning and memory could then account for some of the variance found in musical emotion decoding abilities across listeners.

Studies investigating the mechanisms involved in emotion recognition have mainly focused on features of the musical structure, such as pitch, mode, melody, and harmony (e.g. Thompson & Robitaille, 1992; Vieillard et al., 2008). Less attention, however, has been given to the influence of the performer and his or her individual features in performance, such as articulation and timing (Gabrielsson & Juslin, 1996). Different emotions might be perceived or induced by different performances of the same musical piece (Daly et al., 2015; Yang et al., 2008). It is therefore not just the musical structure that is important for emotion perception but also the way a piece of music is performed.

Decoding factors

While the two factors of the functionalist perspective provide a theoretical basis for emotion decoding in music, the process can still be challenging to study empirically. The subjectivity of music listening and potentially many individual differences can play a role in how accurate listeners are in identifying the intended emotion. This section briefly discusses several key factors that have been investigated previously as possible mediators or moderators in the process of emotion decoding in music.

Emotional intelligence

Trait emotional intelligence (EI), also known as emotional self-efficacy, is associated with personality and refers to the self-perception of emotional abilities (Petrides & Furnham, 2003). Interestingly, Resnicow, Salovey, and Repp (2004) showed that people’s ability to decode emotional expression (happiness, sadness, anger, and fearfulness) in classical piano performance was correlated with their EI ($r = .54$). Such a

correlation suggests that EI could affect, and possibly even predict, emotion decoding abilities in music performance.

Musical training

Many studies have confirmed the positive effect of musical training on memory for music (Cohen, Evans, Horowitz, & Wolfe, 2011), verbal memory (Chan, Ho, & Cheung, 1998) and IQ (Schellenberg, 2011; Schellenberg & Mankarious, 2012), among other cognitive skills. Unfortunately, a consensus on the effect of musical training on musical emotion decoding abilities has yet to be reached, with some studies finding no effect of musical training (e.g. Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005; Campbell, 1942; Juslin, 1997) and other studies finding an effect of musical training on musical emotion decoding accuracy (e.g. Juslin, 2013; Park et al., 2014; Schellenberg & Mankarious, 2012). Due to the lack of consistency in the results, musical training should be investigated as a possible predictor for emotion decoding abilities in music.

Emotional subscale of Gold-MSI

As a way of measuring active involvement in music in its various different forms, Müllensiefen, Gingras, Musil, and Stewart (2014) created the Goldsmiths Musical Sophistication Index (henceforth Gold-MSI). The emotional subscale of the Gold-MSI self-report inventory assesses the degree of expertise when individuals use music to comprehend and alter emotional and mood states and how they process music emotionally. As this scale considers behaviours related to emotional responses to music, it can be considered a potential factor for predicting emotion decoding abilities in music.

Emotional contagion

Emotional contagion refers to the internal mimicking of emotional expression (Mayer, Roberts, & Barsade, 2008; Mayer & Salovey, 1997; Salovey & Mayer, 1990). This psychological mechanism can occur in music but also in other expressive art forms (Egermann & McAdams, 2013). Emotional contagion is an unconscious automatic mechanism of mimicking others' expressions that affect one's own state (Decety & Jackson, 2004; Egermann & McAdams, 2013; Juslin & Västfjäll, 2008; Preston & de Waal,

2002). Previous studies have found a relationship of emotional contagion and emotional reactions to music (Egermann & McAdams, 2013), as well as with the ability to predict and detect emotional reactions in others (Mohn, Argstatter, & Wilker, 2010). We are therefore expecting that individuals with high emotional contagion scores should perform better on tests of emotion decoding ability.

Perhaps the most cited paper investigating how musical performers express and communicate various basic emotion and the accuracy of listeners in decoding the expression was published by Gabrielsson and Juslin (1996). The researchers asked three performers – a flutist, violinist, and vocalist – to record three melodies with seven different emotional expressions. The three melodies used were *Te Deum* by Charpentier, a Swedish folk song, and a novel melody that had been composed specifically for the study. Musicians were instructed to perform the melodies to express happiness, sadness, anger, tenderness, solemnity, fear (only used in their second study), and without expression. Three listening experiments were conducted (one per melody) during which a total of 56 musicians and music psychology students judged the performances in regard to each emotional expression using Likert scales. Comparisons between mean ratings of the intended emotion and all other emotions revealed high decoding accuracies across listeners. However, Gabrielsson and Juslin only analysed the results of the novel melody due to insufficient sample sizes for the other two melodies. Furthermore, they found the singing voice to be far less expressive than the flute and violin and thus excluded ratings of the vocal recordings from their analyses. Additionally, they found that females' mean ratings displayed a trend towards greater accuracy than males' mean ratings, but this difference was not statistically significant. Confusions were found overall between the emotional expressions of sadness and tenderness. Instrument-specific confusions between emotions were also found; the flutist's angry recording was judged as happy or expressionless and the violinist's solemn recording was viewed as angry. In addition, the authors analysed acoustical properties of the recordings and found similar results to earlier studies (Gabrielsson, 1995). For example, happiness and anger were expressed by high sound level, sharp contrast between long and short notes, and bright or harsh timbre. Sadness, fear, and tenderness were expressed by means of large deviations in timing and low sound

level. Solemnity was expressed by small deviations in timing and sharp tone onsets. The performance with no intended expression showed almost no deviation in timing (Gabrielsson & Juslin, 1996).

Considering the important implications of Gabrielsson and Juslin's (1996) study for subsequent research on musical emotion perception and that the paper had been cited 560 times (by April 2018), the present study sought to replicate and extend Gabrielsson and Juslin's (1996) study. A replication would serve to strengthen the reliability of their general findings (Frieler et al., 2013) and shed further light on some of their detailed results that can be considered surprising considering later studies. Despite the lack of expressiveness in the vocalist recordings in the original study, emotion decoding has been shown to be similar between music and voice (for review, see Juslin & Laukka, 2003) which could imply that decoding accuracy of vocal expressions of emotion may have been higher in the original study had there been a more expressive vocalist (see also Weiss, Trehub, & Schellenberg, 2012). Additionally, while only flute, violin, and voice were used in their study, several subsequent studies investigating emotions in music have used piano (e.g. Dalla Bella, Peretz, Rousseau, & Gosselin, 2001; Fritz et al., 2009; Nair, Large, Steinberg, & Kelso, 2002; Sloboda & Lehmann, 2001) and some have found successful emotion decoding using piano performances (Bhatara, Tirovolas, Duan, Levy, & Levitin, 2011; Juslin, Friberg, & Bresin, 2002; Resnicow et al., 2004). Furthermore, in their meta-analysis, Juslin and Laukka (2003) reported the piano to be the third most-studied instrument in music and emotions studies (behind voice and guitar). Therefore, one of the objectives of the present study was to add piano recordings and observe differences in decoding accuracy across instruments. As in the original study, it was hypothesised that there would be overall high decoding accuracies. Because several experiments have found highest accuracies for happy, sad, and angry expressions (e.g. Juslin & Laukka, 2003; Mohn et al., 2010; Peretz, Gagnon, & Bouchard, 1998; Robazza, Macaluso, & D'Urso, 1994), similar patterns were expected in the present study. One of the goals of the replication study was to assess the stability of the findings regarding confusions across and within instruments as reported in Gabrielsson and Juslin's (1996) paper. Similar to Gabrielsson and Juslin's (1996) paper this study also conducted analyses on the acoustical properties of the recordings.

For the extension part of this study, the primary aim was to identify individual differences factors that distinguish between listeners in terms of their emotion decoding accuracy. Considering previous research on individual differences in this area (Daly et al., 2015; Yang et al., 2008), it was hypothesised that musical training, emotional engagement with music, general emotional intelligence, emotional contagion, and gender can potentially be relevant predictors for emotion decoding abilities. The present study was implemented as a multi-lab study with about half of the participants being tested in a controlled lab setting while the other half were tested over the internet.

Method

Design

The current study employed a mixed measures design. The between-subjects variable was the melody each participant was exposed to and the within-subjects variables were the four different instruments (flute, piano, violin, and voice) and the seven emotional expressions as performed by the musicians ("angry", "expressionless", "fearful", "happy", "sad", "solemn", and "tender"). The dependent variables were agreement scales (0–10) for how well each emotion reflected the musical excerpt. The independent variables (predictors) comprised the experimental factors of *Melody*, *Instrument*, and *Intended Expression*. In addition, individual differences measures for *Musical Training* and the *Emotional Subscale* from the Gold-MSI were collected from participants as well as *Emotional Intelligence*, and *Emotional Contagion*.

Participants

In total, 319 participants (103 males, 213 females, 3 other) with varying degrees of musical backgrounds completed the study. The median score for the Gold-MSI Musical Training Scale was 31, which is slightly higher than the median Musical Training score of the general population (median = 27) as reported in Müllensiefen et al. (2014). Participant age ranged from 19 to 69 years ($M_{\text{age}} = 30.3$, $SD = 14.1$). The age distribution was skewed to the right, with 75% of the sample being younger than 33 years. Over half of the participants ($N = 185$, 58%) completed the study in a controlled lab setting, while the others ($N = 134$, 42%) completed the study elsewhere,

unsupervised. Testing was carried out by five labs: Goldsmiths University of London, Max Planck Institute for Empirical Aesthetics (Frankfurt, Germany), Universität Bremen, Katholische Universität Eichstätt-Ingolstadt, and Louisiana State University (Baton Rouge, U.S.A.). Our primary participant recruitment strategy was to gather data from as many participants as possible within an externally determined timeframe to achieve maximal power for the replication. The size of the current sample ($N=319$) is almost 10 times larger than the sample in the original study ($N=34$). In addition, a post-hoc power analysis ($\alpha=.05$, $\text{power}=.8$) indicated that a sample size of about 50 participants tested on the current within-participants design would be sufficient for detecting even the smaller effects that were discussed as interesting by Gabrielsson and Juslin (1996) but did not reach the common level of significance in the original study. This project was approved by the ethical committee of the Psychology Department at Goldsmiths, University of London.

Materials

Four classically trained musicians – a flutist, a pianist, a violinist, and a vocalist – recorded the three melodies used in Gabrielsson and Juslin (1996): *Te Deum* by Charpentier (Melody A), a Swedish folk song (Melody B), and a novel melody specifically composed for their study (Melody C; see Appendix 1). Each musician had at least one university degree in performance on their instrument. The musicians were instructed to perform each melody in seven emotional expressions. Without changing the pitches of the melody, musicians were free to alter other aspects of the performance in any way necessary to communicate the emotions to hypothetical audiences. The vocalist was additionally required to perform all excerpts using a consistent syllable of her choosing; she performed all notes using the vowel [a]. The musicians recorded each emotional rendition twice, both times as similarly as possible. The flutist, pianist, and vocalist were recorded in professional recording studios on a university campus; the violinist used her own recording equipment at her home studio. The recordings were completed and mastered at the Goldsmiths Electronic Music Studios and were edited by a professional sound engineer using Logic Pro 9 to equalise the sound level for all recordings and create a homogeneous set of stimuli. The primary investigators and a research assistant selected the most technically-

accurate version from the two recordings from each musician for use in the present study. Thus, there were 84 excerpts used as stimuli in the present study (i.e. four musicians playing three melodies in seven different emotional expressions).

Individual difference measures

Emotional intelligence

Emotional Intelligence scores were collected through the Trait Emotional Intelligence Questionnaire Short Form (TEIQue-SF; Petrides, 2009). This consisted of a 30-item list of questions to be rated on a 1–7 scale, 1 being “Completely Disagree” and 7 being “Completely Agree”. The TEIQue-SF assessed emotional intelligence as a personality trait by means of self-report questions (e.g. “I can deal effectively with people”, “I usually find it difficult to regulate my emotions”, “I often pause and think about my feelings”).

Emotional contagion

Emotional contagion was determined by using the Emotional Contagion Scale (Doherty, 1997). This 15-item scale was developed to measure one’s susceptibility to others’ emotions. Answers were given on a scale of 1–5, 1 being “Never” and 5 being “Always” (e.g. “I cry at sad movies”, “I melt when the one I love holds me close”). Even though the scale has not been used in music emotion research so far, it has been referred to in numerous studies that investigate the possible relationship between music emotion and contagion (e.g. Egermann & McAdams, 2013; Juslin & Västfjäll, 2008; Vuoskoski & Eerola, 2012).

Musical training and emotional musical sophistication (Gold-MSI)

Additionally, musical training as well as emotional engagement and expertise with music were assessed using the self-report questionnaire portion of the Gold-MSI (Müllensiefen et al., 2014). The Musical Training subscale had seven items, which measured the amount of musical training and practice and the amount of self-assessed musicianship. The Emotions subscale of the Gold-MSI consisted of six items that combined questions about behaviour related to emotional responses to music (e.g. “I am able to talk about the emotions that a piece of music evokes in me”, “I sometimes choose music that can trigger shivers down my spine”) (Müllensiefen et al., 2014). For both Gold-MSI scales responses were given on 7-point Likert scales.

Procedure

All testing was completed using the internet-based survey software Qualtrics. Participants were presented with instructions for the main task of the study, as used in the original study. Participants were randomly assigned to one of the three melody conditions (28 trials) and listened to all seven expressions by one musician in a randomised order before hearing the next musician's recordings in a randomised order, and so on until participants heard all four instruments. The order in which participants heard the instruments was also randomised. Participants rated each of the 28 excerpts in terms of all seven emotional adjectives using Likert scales from 0 to 10. Participants were allowed to listen to the recordings as many times as they wished regardless of whether they participated in the lab or through the online survey. The duration of the test session was approximately 30–40 minutes.

Results

Replication

For the sake of replication, results reported here align directly with those reported in Gabrielsson and Juslin (1996) with the exception of including the piano as an additional instrument into the analysis. Results will first be reported using Melody C, as in Gabrielsson and Juslin's (1996) study. Before analysis, data was screened for responses in which participants gave the same response throughout the test. This resulted in one participant being excluded from all subsequent analysis. One hundred and two participants listened to Melody C. Mean ratings for each emotion across all instruments and with respect to each intended expression (i.e. target emotion) are given in Table 1. To assess the significance of differences between ratings for the different emotions, the data were modelled with the "lme4" package (Bates, Maechler, Bolker, & Walker, 2015) for R using mixed effects models with Rating as the dependent variable, Type of Emotion Rated as independent variables, and Participant as random factor. We ran seven separate mixed effects models, one for each Intended Expression. The difference between the ratings for the target emotion and each of the other six emotions was represented by the model coefficient for each of the rated emotions with reference to the target emotion. The p-value associated with the coefficient estimate was used to indicate the

significance for the difference in the ratings for target and comparison emotion. The significance level was Bonferroni-corrected to account for multiple comparisons and the corresponding significance levels are indicated by asterisks in Table 1. Table 1 therefore summarises the mean emotion ratings across all four instruments as descriptive statistics and also indicates significant differences between the target emotion and each of the other six emotions derived from seven mixed effects models. Subsequently, we ran another seven separate mixed effects model (one for each Intended Expression) which included Instrument and the interaction between Instrument and Type of Emotion Rated as additional fixed effects. In all seven models the interaction between Type of Emotion and Instrument was significant according to an ANOVA Wald χ^2 -test (type III; χ^2 values ranging from 71 to 334, all $dfs = 18$, all p -values $< .001$). Therefore, we computed further separate mixed effects models for each of the four instruments and each Intended Expression (28 models altogether). Summaries of the corresponding instrument-wise mixed effects regression models are given in Tables A1–A4 in Appendix 2.

For an easier understanding the confusion matrix from Table 1 is visualised in Figure 1. The graph shows that for 4 out of 7 emotions (57%), the target emotion received the highest average ratings (angry, happy, sad, tender), while for three emotions this was not the case. In fact, for fearful and solemn as target emotions, tender and sad received the highest average ratings.

However, the patterns of confusions differ noticeably by instrument as can be seen from the confusion matrices in Tables A1–A4 in Appendix 2. For violin, the target emotion received the highest ratings for 4 out of 7 emotions (57%) and for voice 5 out of 7 emotions (71%) obtained highest ratings when they served as target emotion. In contrast for flute only 1 out of 7 (14%) and for piano only 3 out of 7 emotions (43%) were rated highest on average when used as target emotion. A direct comparison shows that the number of times the target emotion received the highest average ratings are generally similar to the results that Gabrielsson and Juslin (1996) reported in their Table 1. The average decoding accuracy aggregated across all participants in our study was lower for violin and flute with 4 out of 6 for violin (67%) and 1 out of 6 for flute (17%) compared to 5 out of 6 for violin (83%) and 3 out of 6 for flute (50%) in the original study.

Table 1. Model-based significant differences and mean emotion ratings (columns) by intended expression (rows) across all instruments.

IntExpr	Angry	Fearful	Happy	NoExpr	Sad	Solemn	Tender
Angry_Int $R^2 = .19$	5.26	1.80***	2.33***	1.81***	1.83***	2.18***	1.20***
Fearful_Int $R^2 = .25$	0.90***	3.77	1.56***	1.76***	4.92***	3.0***	5.01***
Happy_Int $R^2 = .08$	1.77***	1.81***	4.05	2.01***	2.00***	2.75***	2.59***
NoExpr_Int $R^2 = .06$	1.53***	1.83***	1.99***	3.04	3.29 ($p = .14$)	2.90 ($p = .39$)	3.24 ($p = .24$)
Sad_Int $R^2 = .45$	0.64***	2.84***	1.02***	1.48***	6.50	4.15***	5.67***
Solemn_Int $R^2 = .19$	1.42***	2.40***	1.44***	2.10***	4.72***	3.88	3.81 ($p = .66$)
Tender_Int $R^2 = .32$	0.80***	2.80***	1.58***	1.76***	5.41 ($p = .24$)	3.93	5.60

Notes: The significance of the differences between the ratings for the intended expression and all other emotion ratings was assessed by a mixed effects model for each of the seven intended emotional expressions (see description in text). The effect size of each model (i.e. effect of type of emotion rated) was computed as marginal R^2 values for mixed effect models using the approach suggested by Nakagawa and Schielzeth (2013). Degrees of freedom for all significance tests are 2775. Only p -values for non-significant effects are reported. Reported significance levels are Bonferroni corrected. Significance levels are coded as: *** $p < .001$, ** $p < .01$, * $p < .05$.

To obtain a more detailed comparison between the two studies considering the full pattern of emotion ratings, we correlated the average ratings for violin and flute as reported by Gabrielsson and Juslin (1996, p. 76) with the average ratings for the same instruments obtained in our study (Tables A1 and A4 in Appendix 2). Table 2 reports Spearman's rank correlation coefficients that merely reflect differences in the order of average emotion ratings. In addition, Table 2 also gives Pearson's correlation coefficients that reflect the distance of each average rating from the sample mean of all average emotion ratings (for each target emotion).

For violin, the average Pearson's correlation coefficients across all six target emotions was $r = .73$ and for flute $r = .38$. The corresponding rank correlation coefficients were $\rho = .57$ for violin and $\rho = .32$ for flute. The highest agreement between the original study and

the replication of $r = .94$ ($p = .005$) was obtained for Sad and Flute. The lowest agreement between the two studies resulted for No expression and Flute with $r = -.66$ ($p = .174$).

While it is certainly possible to aggregate participants' emotion ratings by averaging, it is also instructive to inspect the distribution of ratings on each emotion scale and for each target emotion. The distributions (of both violin and flute ratings) are depicted as density plots in Figure 2. The figure shows that even for the cases where the target emotion receives by far the highest average rating (i.e. Angry and Sad as target emotions), the distribution of ratings on the corresponding emotion scale is very broad and does not show a clear central tendency.

Extension

Binary response accuracy

As part of the extension of the study, response accuracy was computed as an additional dependent variable at the level of the individual trial by assigning '1' for a correct response, i.e. if the participant had assigned the highest rating for the intended expression (target emotion). '0' was assigned in all other cases. Ratings that tied, meaning the highest score was assigned to the intended emotion but also to another emotion, were considered incorrect. Accuracy scores were computed for all intended expressions and all melodies. The overall accuracy score averaged across all participants was 30.6% (SD = 11.9%; median = 28.6%) which is about twice as high as chance level on the 7-alternative forced choice task at 14.3%. However, accuracy scores for individual participants range from 3.6% to 64.3% with about 5% of participants scoring at chance level or below.

Table 2. Pearson's correlation coefficients (r) and Spearman's rank correlation coefficients (ρ) as well as associated significance values for the correlation between mean emotion ratings in Gabrielsson and Juslin's original (1996, p. 76) and the current study (Tables A1 and A4 in Appendix 2).

Instrument	Intended expression	ρ	p	r	p
Flute	Angry	.00	1.000	.09	.864
	Happy	.09	.919	.00	.994
	No expression	-.66	.175	-.44	.386
	Sad	1.00	.003**	.94	.005**
	Solemn	.66	.175	.87	.024*
	Tender	.83	.058	.81	.051
	Mean	.32		.38	
Violin	Angry	.83	.058	.97	.001**
	Happy	.49	.321	.71	.113
	No expression	-.14	.803	.03	.950
	Sad	.66	.175	.84	.037*
	Solemn	.77	.103	.91	.012*
	Tender	.83	.058	.93	.007**
	Mean	.57		.73	

* $p < .05$.

** $p < .01$.

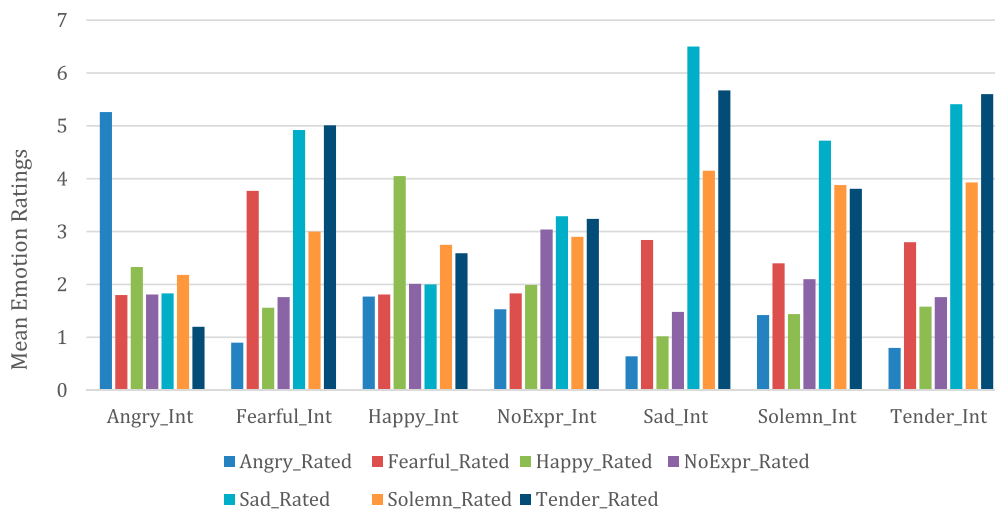


Figure 1. Mean rating of emotions by intended expression across all instruments.

Individual differences measures

Descriptive statistics for the individual differences measures from the questionnaire data as well as their association with decoding accuracy calculated from the binarised responses are given in Table 3.

Associations range from $r = .075$ for Musical Training to $r = .009$ for the Emotional Contagion scale.

Men had a mean decoding accuracy of 0.299 (SD = 0.458) while women showed slightly higher performance (mean accuracy = 0.309, SD = 0.462). However, this difference was not significant according to a t -test ($t(188) = -0.698, p = .486$). Accordingly, Cohen's d indicated a very small effect of gender ($d = .09$).

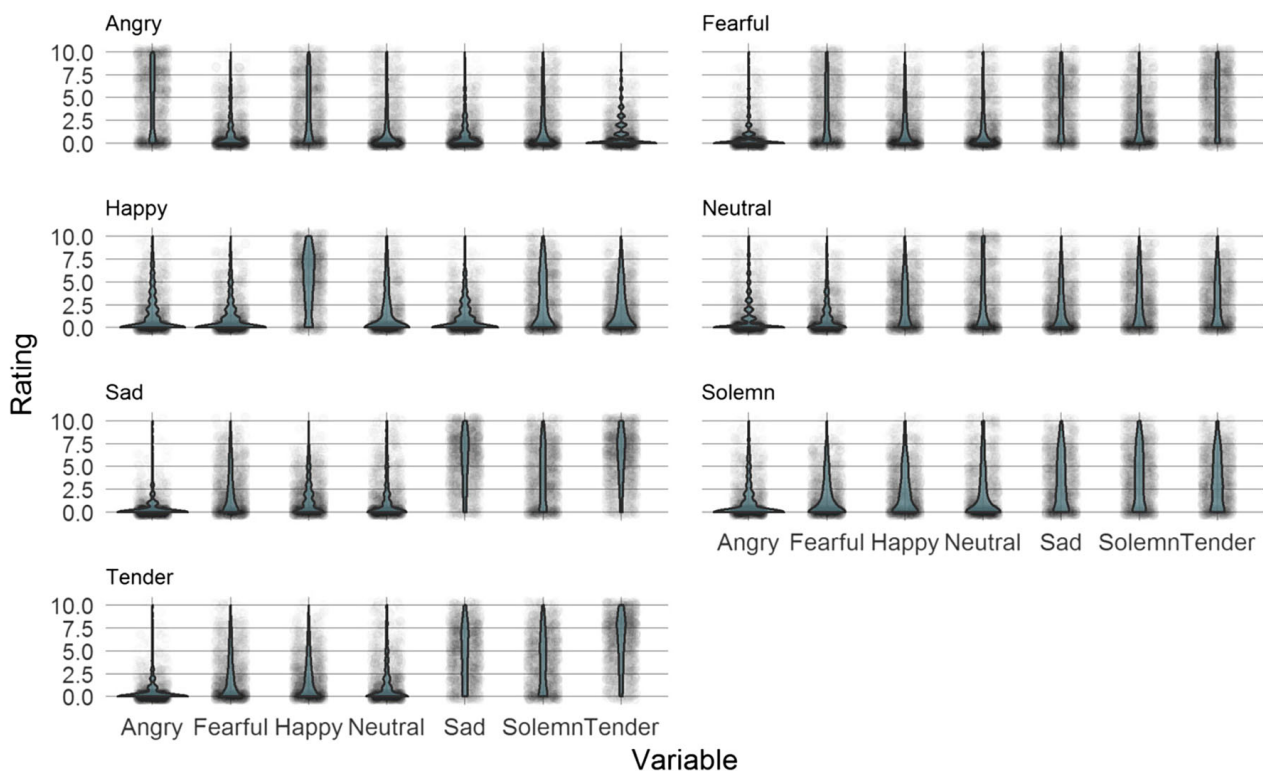


Figure 2. Density plot of emotion ratings by intended expression (separate panels) across all instruments and melodies combining a mirrored density plot with jittered point values. Density estimates were obtained via the R function `density()` using Gaussian kernels.

Table 3. Means, standard deviations and association with binary response accuracy scores for the individual differences measures.

Factor	AM	SD	$r(p)$
Musical training (Gold-MSI, scale range: 7–49)	29.1	10.5	.075 (<.001)
Emotional subscale (Gold-MSI, scale range: 7–42)	33.9	4.9	.042 (<.001)
Emotional intelligence (TEIQue-sf, scale range: 30–210)	154.5	19.3	.021 (.04)
Emotional contagion (ECS scale range: 15–75)	51.4	8.7	.009 (.4)
Gender	–	–.01	–.01 (.33)

Notes: Higher scores generally indicate higher levels of the assessed trait. Pearson correlations were calculated for all continuous variables and the point-biserial correlation was calculated for Gender. AM = arithmetic mean, SD = standard deviation, $r(p)$ = Pearson's correlation coefficient and corresponding p -value.

Table 4 shows descriptive statistics of emotion decoding accuracy scores across the five different labs and between lab and online participants. A linear model with emotion decoding accuracy score as dependent variable showed that neither test location ($F(4) = 1.164$, $p = .33$) nor test environment ($F(1) = 0.004$, $p = .95$) affected performance accuracy. Therefore, data were collapsed across these two factors for the subsequent analyses.

Joint modelling of experimental and individual differences factors

To investigate the relationship between the binary emotion decoding accuracy scores and the individual factors, the “lme4” package (Bates et al., 2015) for R was used to fit a series of generalised linear mixed effects models, starting with an initial null model with fixed effects for the three experimental factors, *Melody*, *Instrument*, and *Intended Expression* and *Participant* as random effect. Further predictors were added to the null model in a hierarchical step-by-step fashion, starting with the individual differences measures that were most closely associated with emotion decoding accuracy according to Pearson's

Table 4. Means and standard deviations for emotion decoding accuracy scores across the five labs (test location) and between in-lab and online participants (test environment).

Lab	N (in-lab)	N (online)	Mean	SD
U Bremen	7	11	.33	.12
MPI	67	28	.29	.11
Goldsmiths	31	76	.30	.11
KUE	42	7	.33	.12
LSU	36	10	.32	.14
Overall lab	183	–	.31	.12
Overall online	–	132	.30	.12

correlations (i.e. Musical Training, Emotions Subscale from the Gold-MSI, Emotional Intelligence, Gender, Emotional Contagion). Likelihood ratio tests were then used to compare each model to the next complex model (i.e. containing one more predictor) in consecutive order. Results indicated that the second model differed significantly from its predecessor ($\chi^2(1) = 28.99$, $p < .001$), but did not differ significantly from the third model ($\chi^2(1) = 0.2$, $p = .65$). Thus, only Musical Training and none of the emotion-related self-report scales nor *Gender* made a significant contribution to explaining emotion decoding accuracy. This result was confirmed by an alternative approach to model selection based on computing the corrected Akaike Information Criterion for all candidate models (see Long, 2012). The model only including Musical Training had the smallest AICc value (AICc = 10221) compared to the next model which also included the emotions subscale from the Gold-MSI (AICc = 10223) and any of the models including three or more individual differences measures and the null model (AICcs > 10224). Hence, the model only including Musical Training as individual difference measure was selected as the final model. The overall model fit was $R^2_{\text{marginal}} = 0.11$ and $R^2_{\text{conditional}} = 0.15$. The effects of the four model predictors are visualised in Figure 3. The corresponding table of regression coefficients (Table A5) is given in Appendix 3.

Acoustical modelling

In order to obtain an understanding of the musical cues used by the performers to convey the intended emotions, we extracted acoustical features from the audio recordings, using the MIRToolbox for MATLAB (Lartillot & Toivainen, 2007). We chose this approach as a numerical alternative to the more qualitative assessment used by Gabrielsson and Juslin which has become a standard tool for modelling music emotion perception over recent years. To this end, we extracted 32 acoustical (see Appendix 4) features plus duration (as an indicator for tempo) that have previously proved to be rather successful in prediction of emotional expression for complex music (Lange & Frieler, 2018).

All features were melody and instrument-wise z -transformed to exclude influences from melody characteristics and instrument timbre. All z -values are hence relative to the mean of the 12 classes (3 melodies times 4 instruments). Features were

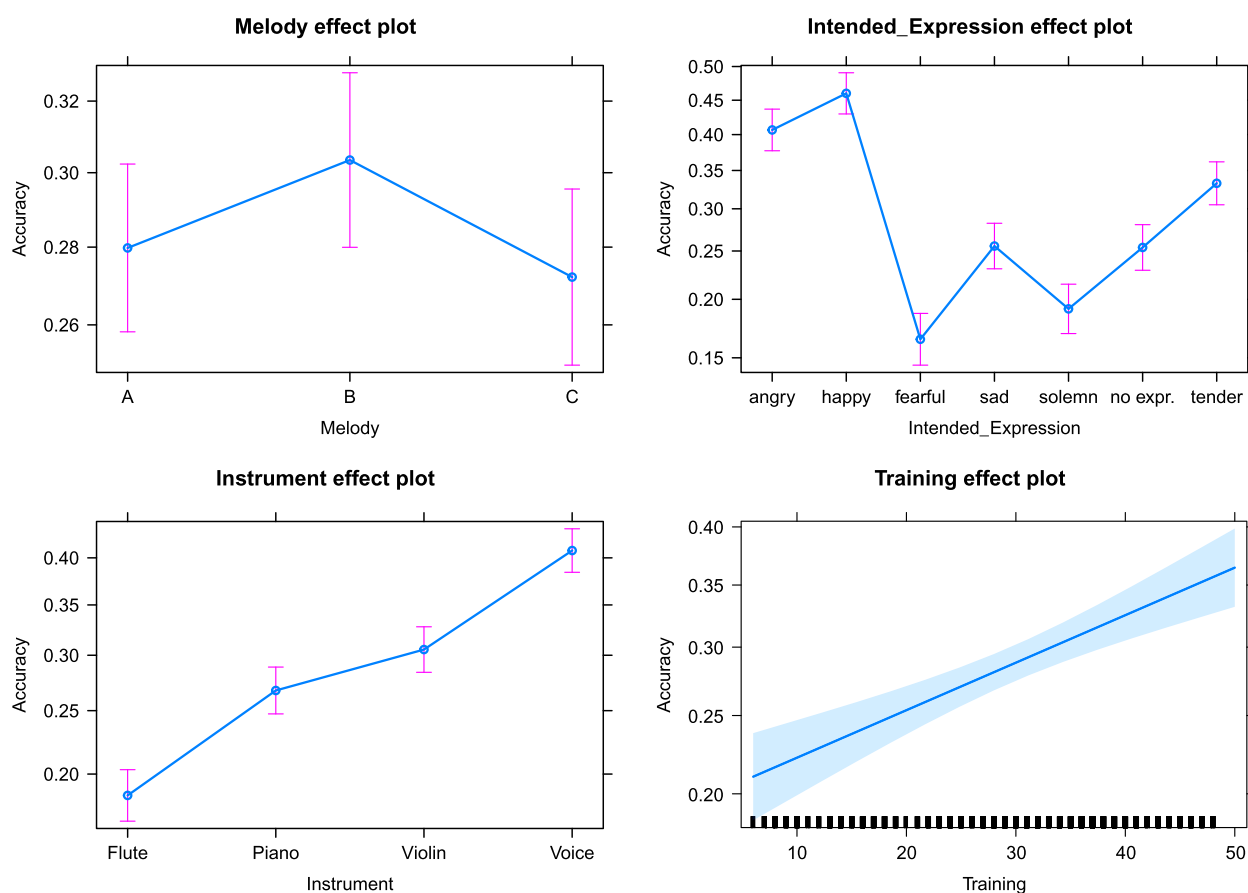


Figure 3. Relationships between the four fixed effect predictors (Melody, Intended Expression, Instrument, and Musical Training) and emotion decoding accuracy from the final mixed effects model. The model also included a random effect for Participant (not shown).

screened for high inter-feature correlations of $|r| > .9$, which are mostly due to the similar mathematical construction of several features. This left a final set of 21 features. No PCA or any other dimension reduction method was employed at this stage to in order to keep the interpretation of the features simple. In a next step, we used random forests (Breiman, 2001) to classify expressed emotion with these feature values. Random Forests are among the most powerful and easy to interpret classification

algorithms and conditional random forest implementation from the R package party (Hothorn, Hornik, & Zeileis, 2006) has been shown to be fairly robust against collinearity (Strobl, Boulesteix, Zeileis, & Hothorn, 2006). The confusion matrix of the model-based classifications is shown in Table 5. Overall classification accuracy was 46% (baseline 14%). The highest classification accuracy was found for angry and fearful, with the lowest accuracy rate for tender, happy, and solemn.

Table 5. Confusion matrix for random forest classification with 1000 trees of emotional expression with 21 acoustical features.

	Angry	Fearful	Happy	No expression	Sad	Solemn	Tender
Angry	.62	.00	.38	.00	.00	.00	.00
Fearful	.00	.77	.00	.00	.00	.08	.15
Happy	.17	.17	.25	.42	.00	.00	.00
No expression	.00	.00	.27	.45	.00	.27	.00
Sad	.00	.00	.00	.00	.50	.08	.42
Solemn	.00	.00	.00	.18	.09	.36	.36
Tender	.00	.00	.00	.00	.56	.33	.11

Notes: Columns: intended emotions, rows: identified emotions. All classification rates row-wise normalised; maximum values in bold.

The six most important features according to the variable importance measure from the random forest model were: *duration* (tempo), *rms_mean* (intensity), *brightness_mean* (Brightness), *flatness_mean* (noisiness), *roughness_mean* (dissonance, roughness), *spec_entropy_mean* (timbre complexity). The distributions (densities) of these features according to intended expression provide valuable insights (see Figure 4). Interestingly, the distributions are sometimes multi-modal and/or fairly broad. This suggests divergences in the interpretations of the different performers or differences between melodies.

Angry is the most extreme expression, and is characterised by high roughness, brightness, loudness, and timbral complexity, as well as the fastest tempos (shortest durations). Happy has similar characteristics, hence the confusion of happy and angry (see Table 5), but is less extreme. Sad expression is conveyed by slow tempos, low dissonance/roughness, brightness and soft intensities. Tenderness is nearly indistinguishable from sadness on these features, which can be already seen in the confusion matrix. Fear is similar to sadness, but with greater timbral complexity and faster tempos. Solemn is also very similar to sadness but has slightly faster tempos,

higher intensities and more roughness. Finally, neutral expression is very similar to happy and solemn (faster than solemn but slower than happy) and with less timbral complexity than happy but more than solemn.

Discussion

The present study intended to replicate and extend Gabrielsson and Juslin's (1996) study. The melodies, instructions to performers, instructions to listeners, emotional expressions, and reporting scales were identical to those used in the original study. However, as the current study was presented as an internet-based survey, it may not be appropriate to consider the present study as an "exact" replication; nonetheless, recognising the need for replication in music psychology (Frieler et al., 2013), the growing tendency to collect data online using internet-based surveys (Reips, 2012), the successful use of an internet-based survey for a music emotion recognition study (Egermann & McAdams, 2013), and the utilisation of identical materials and procedures as the original study arguably warrants consideration of the present study as a replication. Results of the present

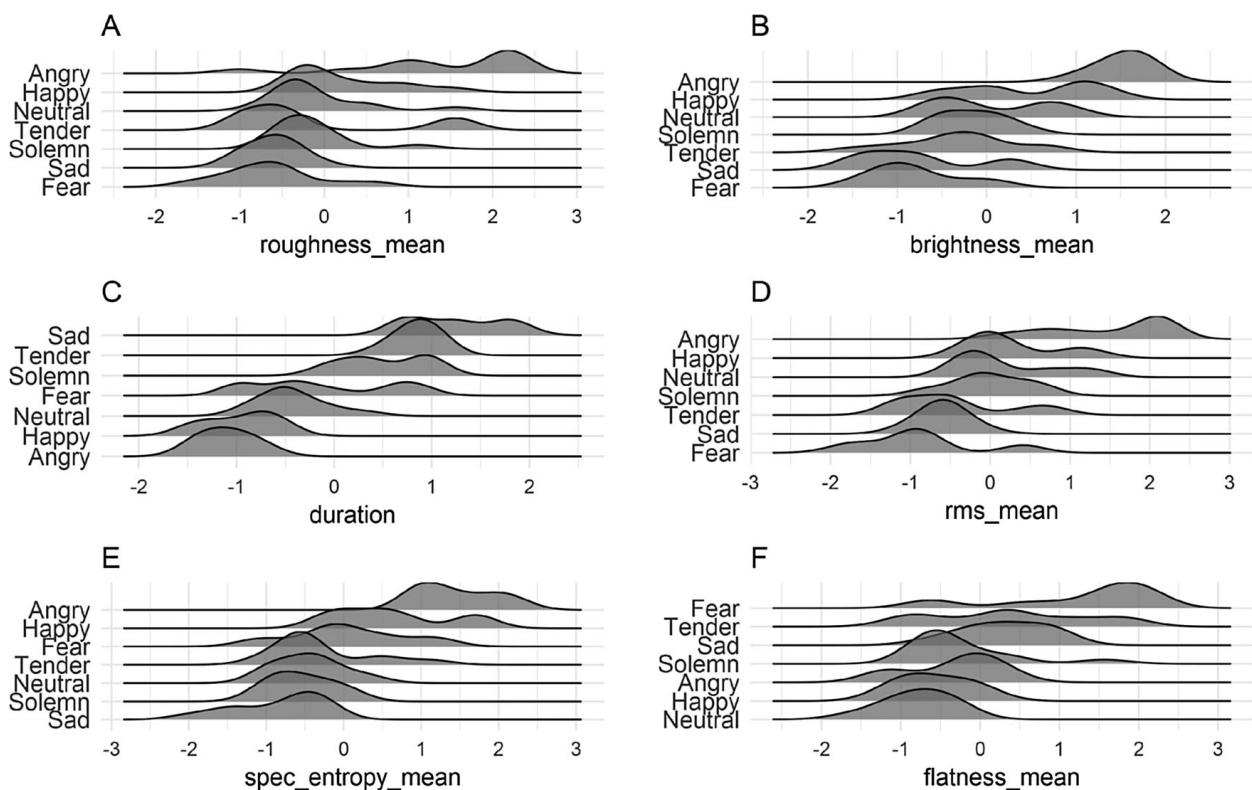


Figure 4. Ridgeline plots of the six most important acoustical features with respect to intended emotion pooled over all three melodies. The intended emotions on the y-axis are ordered according mean values of the feature distribution. The x-axis shows z-values of the features.

study were divided into two parts: the first considering average ratings for melody C matching Gabrielsson and Juslin's (1996) results; the second part used data at the individual trial level and from all melodies, analysing the effect of experimental as well as individual difference factors (as part of the extension). The results from the first part support the original study's findings of overall high decoding accuracy with 4 out of 7 emotions (57%) receiving highest average ratings when the emotion served as the expression intended by the performers. As in the original study, confusions were evident between sad excerpts and solemn or tender. Additionally, fearful excerpts performed by violin were confused and solemn excerpts performed by piano were confused. Tender, solemn and no expression excerpts performed by flute were poorly recognised.

While Gabrielsson and Juslin (1996) found the voice to be least expressive in their study, the present study showed that the voice significantly outperformed piano and flute in conveying anger and tenderness. Additionally, the voice and violin were the only instruments to achieve a mean rating of close to or greater than '7' for a congruent emotion (for tenderness and sadness). Importantly, the voice used in the original study was male and our vocalist was female. It is possible that the vocalist's gender can affect judgments of emotionality (Scherer, Banse, & Wallbott, 2001), which may explain the discrepancy between expressiveness judgments in the original paper compared to the present study. Thus, future research using a larger sample of vocal performers from both genders is needed to verify this claim. Nonetheless, the finding that the voice was more expressive than the flute and violin supports the numerous studies in the field that have studied music and emotions using the voice (for review, see Juslin & Laukka, 2003). In fact, in their meta-analysis, Juslin and Laukka (2003) found singing to be the most frequently studied instrument in research on music and emotions. Furthermore, the finding that the voice and violin were most effective at conveying tenderness may arise from a developed sensitivity to the expression of tenderness following maternal singing and melodic speaking during infancy (see Nakata & Trehub, 2004). Overall, it is largely unsurprising for the voice to be most effective at communicating emotional information as scholars have long considered vocal communication of emotion to be evolutionarily and biologically advantageous (see Scherer, 1995).

Like Gabrielsson and Juslin (1996), the current study found no significant difference in emotion decoding ability for participants' gender. Although some gender differences have been reported in previous music and emotion research, a consensus on gender differences with regards to emotion processing in music has yet to be reached (e.g. Brackett, Rivers, Shiffman, Lerner, & Salovey, 2006; Doherty, Orimoto, Singelis, Hatfield, & Hebb, 1995; Kafetsios, 2004; McRae, Ochsner, Mauss, & Gabrieli, 2008; Petrides, Furnham, & Martin, 2004). Further research is therefore needed to investigate the stereotypical belief that women have stronger emotional processing skills than men (McRae et al., 2008).

The aim of the replication part of the study was to investigate how emotions are identified in music performance. We hypothesised that there would be high overall decoding accuracy. While the results of the original study and replication part of this study were analysed using the mean ratings for the excerpts, for the extension part of the study accuracy scores were acquired by assigning a "1" when the intended expression was rated the highest. This meant that confusion (i.e. highest rating for intended and another emotion) was not considered correct and was assigned a "0". Using this method of scoring showed that there was an overall decoding accuracy of 30.6% which is well above chance level (14%) but substantially lower than the overall accuracy across all participants (57%). This superior performance of the aggregate sample compared to individual participants is in line with the so-called wisdom of the crowd effect known from other perceptual and cognitive tasks (Galton, 1907; Yi, Steyvers, Lee, & Dry, 2012). However, this discrepancy suggests that special care needs to be taken when reporting that 'listeners are generally able to recognise emotions in music'. This seems to be true for aggregate judgements from a large sample but can potentially disguise the fact that most individual participants show rather poor to moderate performances on musical emotion decoding tasks.

Individual factors

Additionally, it was hypothesised that there are individual factors that play a role in decoding accuracy and that these factors can be used as predictors to assess emotional decoding abilities in music performance. The individual factors considered in this study were musical training, the emotion subscale of the

Gold-MSI, EI, emotional contagion, and gender. Using GLMER modelling, a null model only used experimental factors. Adding in Musical Training scores significantly increased the explanatory power of the model. However, no other individual difference measure increased the model fit any further. Hence, only Musical Training seemed to have an effect on decoding accuracy in music. This appears to be contradictory to some of the previous literature that found that training is not important in emotional decoding in performance (e.g. Bigand et al., 2005; Gabrielsson & Juslin, 2002; Juslin, 1997; Vieillard et al., 2008). However, most previous studies only used a fairly coarse indicator of musical training, while the current study was able to make use of the more fine-grained musical training scale from the Gold-MSI. The absence of an effect from the Gold-MSI Emotions subscale might be due to collinearity since the Musical Training and Emotions subscales were correlated by $r = .504$ across participants in the sample. Note that the Emotions subscale itself was significantly correlated with emotion decoding accuracy. The lack of any significant effect from the Emotional Intelligence and Emotional Contagion self-report scales might suggest that decoding intended expressions in music as a cognitive ability is a specific skill that is not necessarily related to general emotion-related traits and abilities. The Emotional Contagion Scale has not yet been used in music research before and therefore its predictive usefulness for measuring emotional contagion has yet to be established.

Two stages were conjectured regarding the process involved in emotion decoding. During the first perceptual stage the listener needs to be able to pick up the relevant and possibly subtle cues in the recording that the performer uses to convey the emotion. This is in congruence with the functionalist perspective of innate mechanism for emotion decoding. Findings from this and other studies found that basic emotions including happy, angry, and sad are easiest to identify in music and while other emotions are generally more difficult (De Gelder & Vroomen, 1996; Gabrielsson & Juslin, 2002; Gabrielsson & Lindström, 1995; Juslin, 1997). However, there were clear interaction effects between different instruments and intended expressions as evident from the confusion matrices for the different instruments and the linear mixed effects models. These interactions provide clear evidence that certain emotions are more easily decoded when conveyed through

specific instruments. Furthermore, there was strong main effect of instrument on accuracy scores showing that expressions conveyed by the voice and violin were easiest to distinguish, while emotion decoding for flute recordings was hardest.

In the second stage of the emotion decoding process the listener needs to interpret the perceived cues and associate them with only one emotion. This is related to cultural differences and social learning, as assumed by the functionalist perspective that the communication of emotions is driven by social interactions. This idea was supported by finding an effect of musical training on decoding accuracy. However, the findings regarding the impact of musical training as well as from other individual difference need to be replicated in the future, ideally using a shorter and more efficient test of emotion decoding accuracy. A more efficient test might use fewer instruments (e.g. piano and voice) and target emotions (e.g. happy, sad, angry, tender) while still ensuring a large range of stimulus difficulty. It seems advisable to compare emotion decoding performance to a larger range of individual difference measures, including perceptual auditory tests, measures of verbal emotion decoding ability, and other musical tasks (e.g. Müllensiefen et al., 2014).

Furthermore, in an attempt to understand the role of performance cues on musical emotion expression and decoding abilities, Gabrielsson and Juslin (1996) analysed musicians' performance cues from tempo, timing, articulation/dynamics, and timbre and found patterns that partly replicated results of the authors' earlier publications (see Gabrielsson & Juslin, 1996). In an effort to fully understand successful decoding of emotions in music, researchers must consider shared cues used by performers when emotions are appropriately understood and variances in cue utilisations between performers when some more accurately convey the emotion than others.

Interestingly, the prediction of emotional expression using 21 acoustical features using random forests showed a higher accuracy (46%) than the average human judgments in our sample (30.6%). A closer scrutiny of the six most important features showed results that are consistent with the characteristics of emotional expression as reported by Gabrielsson and Juslin (1996). Some of the features were multi-modally distributed across intended emotion, which hints at the different usage of acoustical cues across performers (instruments), and might have also affected human emotion decoding

performance. Previous research has revealed that the acoustical limitations and playing methods of different instruments mean that different instruments lend themselves more easily towards different categories of emotional expression (Huron, Anderson, & Shanahan, 2014). To explore the possible interactions between intended expressions, instruments, and performers, it would be necessary to work with a much larger sample of different instruments and different performers (similarly to Study II in Gabrielsson & Juslin, 1996, that compared six different performers playing the same instrument).

A comparative inspection of the confusion matrices arising from this study (i.e. Table 1; p. 76 from Gabrielsson & Juslin, 1996; aggregate ratings from the current study in Table 1 above and the confusion matrix of the statistical model using acoustic features in Table 5 above) provides interesting insights regarding the similarity of emotional categories and potential factors that are underlying these similarity relationships. Firstly, confusions mainly happen between emotion categories that belong to the same arousal level. This is evidenced by the comparatively high average of happy ratings for angry as intended emotion and vice versa as well as by high ratings for solemn and tender when sad is the intended emotion and vice versa. This pattern even holds true for the confusion matrix of the statistical model (Table 5) and confirms that arousal is a primary perceptual factor for the decoding of emotions in music. However, the number of confusions is higher for emotions at the low arousal level (sad, solemn, tender) compared with the high arousal emotions (happy, angry). This higher confusion rate among low arousal emotions is mirrored by the differences in intensity (as measured by the RMS amplitude) as a central acoustic feature that is commonly associated with arousal. As depicted in Figure 4 the RMS amplitude differences between angry and happy are larger than between the three low arousal emotions. In addition, happy and angry differ noticeably in roughness while there is no clear separation between the low arousal emotions. Finally, the confusion matrix given in Table 1 of this study also shows that confusions among the three low level emotions are not symmetric, but that solemn and tender are more frequently confused with sad than the other way around. This asymmetry might be caused by the higher frequency and thus greater cognitive availability of the attribute 'sad' (word frequency rank 2164 in Corpus of Contemporary American

English) relative to 'tender' (frequency rank 4016) and 'solemn' (frequency rank >5000). This is in line with the availability heuristic (Tversky & Kahneman, 1973) that states that when judging under uncertainty the cognitively more available option is often preferred.

Overall, the present study generally supports the findings from Gabrielsson and Juslin (1996). Similarly, high decoding accuracies were found for violin. However, in our study violin and voice were the most expressive instruments (i.e. leading to the greatest decoding accuracies). This finding differs greatly from the findings of Gabrielsson and Juslin (1996) who excluded voice from analyses due to its lack of expressiveness and low decoding accuracy. In summary, the findings reported in the present study support that listeners are generally able to decode musical emotions clearly above chance level for each of the four instruments used in this study. However, decoding abilities are generally far from perfect and differ by emotion type, with basic emotions (i.e. happy, angry, sad) being decoded more easily. Individual differences in emotion decoding ability were partly explained by musical training. However, the moderate effect size of the final model suggests that further factors might still contribute to emotion decoding in music such as extra-musical associations, individual preferences, and personal listening biographies.

Data availability statement

All data and analysis scripts from this study are available through the Open Science Framework: https://osf.io/2eafd/?view_only=fe575c0c93cb4c8c98e8a2edb94e569c.

Acknowledgements

We thank Merel Vercammen, Chris Lee, María Magnúsdóttir, and Sarah Collin for performing the musical excerpts. We thank Björn Þorleifsson for recording and post-processing the recordings of the musical excerpts.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Jessica Akkermans  <http://orcid.org/0000-0002-0619-4374>

References

- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01
- Baumgartner, T., Lutz, K., Schmidt, C. F., & Jäncke, L. (2006). The emotional power of music: How music enhances the feeling of affective pictures. *Brain Research*, 1075(1), 151–164.
- Bhatara, A., Tirovolas, A. K., Duan, L. M., Levy, B., & Levitin, D. J. (2011). Perception of emotional expression in musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3), 921–934. doi:10.1037/a0021922
- Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion*, 19(8), 1113–1139. doi:10.1080/02699930500204250
- Brackett, M. A., Rivers, S. E., Shiffman, S., Lerner, N., & Salovey, P. (2006). Relating emotional abilities to social functioning: A comparison of self-report and performance measures of emotional intelligence. *Journal of Personality and Social Psychology*, 91(4), 780–795. doi:10.1037/0022-3514.91.4.780
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Campbell, I. G. (1942). Basal emotional patterns expressible in music. *The American Journal of Psychology*, 55(1), 1–17. Retrieved from <http://www.jstor.org/stable/10.2307/1417020> \npapers2://publication/uuid/EF1DE689-9D9F-4904-9DED-0E68EEA08A37
- Chan, A. S., Ho, Y. C., & Cheung, M. C. (1998). Music training improves verbal memory. *Nature*, 396(6707), 128. doi:10.1038/24075
- Cohen, M. A., Evans, K. K., Horowitz, T. S., & Wolfe, J. M. (2011). Auditory and visual memory in musicians and nonmusicians. *Psychonomic Bulletin & Review*, 18(3), 586–591. doi:10.1016/j.cmet.2012.08.002
- Dalla Bella, S., Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. *Cognition*, 80(3), 1–10. doi:10.1016/S0010-0277(00)00136-0
- Daly, I., Williams, D., Hallowell, J., Hwang, F., Kirke, A., Malik, A., ... Nasuto, S. J. (2015). Music-induced emotions can be predicted from a combination of brain activity and acoustic features. *Brain and Cognition*, 101, 1–11. doi:10.1016/j.bandc.2015.08.003
- Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, 3. doi:10.1177/1534582304267187
- De Gelder, B., & Vroomen, J. (1996). Categorical perception of emotional speech. *Journal of the Acoustical Society of America*, 100(2818). doi:10.1121/1.416612
- Doherty, R. (1997). The emotional contagion scale: A measure of individual differences. *Journal of Nonverbal Behavior*, 21(2), 131–154. doi:10.1023/A:1024956003661
- Doherty, R. W., Orimoto, L., Singelis, T. M., Hatfield, E., & Hebb, J. (1995). Emotional contagion: Gender and occupational differences. *Psychology of Women Quarterly*. doi:10.1111/j.1471-6402.1995.tb00080.x
- Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1), 18–49. doi:10.1177/0305735610362821
- Egermann, H., & McAdams, S. (2013). Empathy and emotional contagion as a link between recognized and felt emotions in music listening. *Music Perception*, 31(2), 139–156.
- Escoffier, N., Zhong, J., Schirmer, A., & Qui, A. (2013). Emotions in voice and music: Same code, same effect? *Human Brain Mapping*, 34, 1796–1810.
- Frieler, K., Müllensiefen, D., Fischinger, T., Schlemmer, K., Jakubowski, K., & Lothwesen, K. (2013). Replication in music psychology. *Musicae Scientiae*, 17, 265–276. doi:10.1177/1029864913495404
- Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., ... Koelsch, S. (2009). Universal recognition of three basic emotions in music. *Current Biology*, 19(7), 573–576. doi:10.1016/j.cub.2009.02.058
- Gabrielsson, A. (1995). Expressive intention and performance. In *Music and the mind machine* (pp. 35–47). Berlin: Springer.
- Gabrielsson, A., & Juslin, P. N. (1996). Emotional expression in music performance: Between the performers's intention and the listener's experience. *Psychology of Music*, 24, 68–91.
- Gabrielsson, A., & Juslin, P. N. (2003). Emotional expression in music. In R. J. Davidson, K. R. Sherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 503–534). New York, NY: Oxford University Press.
- Gabrielsson, A., & Juslin in James, R. J. D. W. (eds). (2002). *Handbook of affective sciences*. New York, NY: Oxford University Press.
- Gabrielsson, A., & Lindström, E. (1995). Emotional expression in synthesizer and sentograph performance. *Psychomusicology: A Journal of Research in Music Cognition*, 14(1–2), 94.
- Gabrielsson, A., & Lindström, E. (2010). The role of structure in musical expression of emotions. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 367–400). New York, NY: Oxford University Press.
- Galton, F. (1907). Vox populi. *Nature*, 75, 450–451. doi:10.1038/075450a0
- Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 651–674. doi:10.1198/106186006X133933
- Huron, D., Anderson, N., & Shanahan, D. (2014). “You can’t play a sad song on the banjo”: Acoustic factors in the judgment of instrument capacity to convey sadness. *Empirical Musicology Review*, 9(1), 29–41.
- Jäncke, L. (2008). Music, memory and emotion. *Journal of Biology*, 7(6), 21.
- Juslin, P. N. (1997). Emotional communication in music performance: A functionalist perspective and some data. *Music Perception: An Interdisciplinary Journal*, 14(4), 383–418.
- Juslin, P. N. (2013). What does music express? Basic emotions and beyond. *Frontiers in Psychology*, 4, 596. doi:10.3389/fpsyg.2013.00596
- Juslin, P. N., Friberg, A., & Bresin, R. (2002). Toward a computational model of expression in music performance: The GERM model. *Musicae Scientiae*, 5(1 suppl), 63–122. doi:10.1177/102986490200505104
- Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5), 770–814. doi:10.1037/0033-2909.129.5.770

- Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3), 217–238. doi:10.1080/0929821042000317813
- Juslin, P. N., & Sloboda, J. A. (2001). *Music and emotion: Theory and research*. Oxford: Oxford University Press.
- Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *The Behavioral and Brain Sciences*, 31(5), 559–621. doi:10.1017/S0140525X08006079
- Kafetsios, K. (2004). Attachment and emotional intelligence abilities across the life course. *Personality and Individual Differences*, 37(1), 129–145. doi:10.1016/j.paid.2003.08.006
- Lange, E., & Frieler, K. (2018). Challenges and opportunities of predicting musical emotions with perceptual and automated features. *Music Perception*, 38(2), 217–242.
- Lartillot, O., & Toivianen, P. (2007). A Matlab toolbox for musical feature extraction from audio. International conference on digital audio effects (pp. 237–244), Bordeaux, France, September 10–15.
- Long, J. D. (2012). *Longitudinal data analysis for the behavioral sciences using R*. Los Angeles: Sage.
- Mayer, J. D., Roberts, R. D., & Barsade, S. G. (2008). Human abilities: Emotional intelligence. *Annual Review of Psychology*, 59(1), 507–536. doi:10.1146/annurev.psych.59.103006.093646
- Mayer, J. D., & Salovey, P. (1997). What is emotional intelligence? *Emotional Development and Emotional Intelligence*. doi:10.1177/1066480710387486
- McRae, K., Ochsner, K. N., Mauss, I., & Gabrieli, J. (2008). Gender differences in emotion regulation: An fMRI study of cognitive reappraisal. *Group Processes & Intergroup Relations*, 11(2), 143–162. doi:10.1177/1368430207088035
- Mohn, C., Argstatter, H., & Wilker, F.-W. (2010). Perception of six basic emotions in music. *Psychology of Music*, 39(4), 503–517. doi:10.1177/0305735610378183
- 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), doi:10.1371/journal.pone.0089642
- Nair, D. G., Large, E. W., Steinberg, F., & Kelso, J. S. (2002). Perceiving emotion in expressive piano performance: A functional MRI study. In *Proceedings of the 7th international conference on music perception and cognition* (Vol. July, pp. 627–630). Adelaide: Causal Productions.
- Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4, 133–142.
- Nakata, T., & Trehub, S. E. (2004). Infants' responsiveness to maternal speech and singing. *Infant Behavior and Development*, 27(4), 455–464.
- Park, M., Gutyrchik, E., Bao, Y., Zaytseva, Y., Carl, P., Welker, L., ... Meindl, T. (2014). Differences between musicians and non-musicians in neuro-affective processing of sadness and fear expressed in music. *Neuroscience Letters*, 566, 120–124. doi:10.1016/j.neulet.2014.02.041
- Peretz, I., Gagnon, L., & Bouchard, B. (1998). Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. *Cognition*, 68(2), 111–141. doi:10.1016/S0010-0277(98)00043-2
- Petrides, K. V. (2009). Psychometric properties of the trait emotional intelligence questionnaire (TEIQue). *Assessing Emotional Intelligence*, 103–117. doi:10.1007/978-0-387-88370-0
- Petrides, K. V., & Furnham, A. (2003). Trait emotional intelligence: Behavioural validation in two studies of emotion recognition and reactivity to mood induction. *European Journal of Personality*, 17(1), 39–57. doi:10.1002/per.466
- Petrides, K. V., Furnham, A., & Martin, G. N. (2004). Estimates of emotional and psychometric intelligence: Evidence for gender-based stereotypes. *The Journal of Social Psychology*, 144(2), 149–162. doi:10.3200/SOCP.144.2.149-162
- Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *The Behavioral and Brain Sciences*, 25(1), 1–71. doi:10.1017/S0140525X02000018
- Reips, U.-D. (2012). Using the internet to collect data. *APA Handbook of Research Methods in Psychology: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*, 2, 291–310.
- Resnicow, J. E., Salovey, P., & Repp, B. H. (2004). Is recognition of emotion in music performance an aspect of emotional intelligence? *Music Perception*, 22(1), 145–158. doi:10.1525/mp.2004.22.1.145
- Robazza, C., Macaluso, C., & D'Urso, V. (1994). Emotional reactions to music by gender, age, and expertise. *Perceptual and Motor Skills*, 79, 939–944.
- Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition, and Personality*, 9(3), 185–211. doi:10.1016/S0962-1849(05)80058-7
- Schellenberg, E. G. (2011). Music lesson, emotional intelligence, and IQ. *Music Perception*, 29(2), 185–194. doi:10.1525/MP.2011.29.2.185
- Schellenberg, E. G., & Mankarious, M. (2012). Music training and emotion comprehension in childhood. *Emotion*, 12(5), 887–891. doi:10.1037/a0027971
- Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice*, 9(3), 235–248.
- Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology*, 32(1), 76–92.
- Sloboda, J. A., & Lehmann, A. C. (2001). Tracking performance correlates of changes in perceived intensity of emotion during different interpretations of a Chopin piano prelude. *Music Perception*, 19(1), 87–120. doi:10.1525/mp.2001.19.1.87
- Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2006). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8, 25.
- Thompson, W. F., & Robitaille, B. (1992). Can composers express emotions through music? *Empirical Studies of the Arts*, 10(1), 79–89.
- Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science: A Journal of the American Psychological Society / APS*, 11(3), 188–195. doi:10.1111/1467-9280.00240
- Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207–232.
- Viellard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, 22(4), 720–752.

- Vuoskoski, J. K., & Eerola, T. (2012). Can sad music really make you sad? Indirect measures of affective states induced by music and autobiographical memories. *Psychology of Aesthetics, Creativity, and the Arts*, 6(3), 204–213. doi:10.1037/a0026937
- Ward, J. (2006). *The student's guide to cognitive neuroscience*. London: Psychology Press.
- Weiss, M. W., Trehub, S. E., & Schellenberg, E. G. (2012). Something in the way she sings: Enhanced memory for vocal melodies. *Psychological Science*, 23(10), 1074–1078.
- Yang, Y. H., Lin, Y. C., Su, Y. F., & Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions On Audio Speech And Language Processing*, 16(2), 448–457. doi:10.1109/TASL.2007.911513
- Yi, S. K. M., Steyvers, M., Lee, M., & Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive Science*, 36(3), 452–470.

Appendices

Appendix 1. Melodies

Melody A



Melody B



Melody C



Appendix 2. Tables A1–A4**Table A1.** Model-based significant differences and mean rating of emotions (columns) by intended expression (rows) for flute for Melody C.

IntExpr	Angry	Fearful	Happy	NoExpr	Sad	Solemn	Tender
Angry $R^2 = .04$	2.70	2.06 ($p = .06$)	3.54*	2.85*	2.10 ($p = .08$)	2.54 ($p = .65$)	2.35 ($p = .31$)
Fearful $R^2 = .22$.64***	4.16	2.32***	2.84***	4.02 ($p = .62$)	2.41***	4.66 ($p = .11$)
Happy $R^2 = .12$.99***	2.03***	3.38	2.08***	2.85 ($p = .11$)	2.80 ($p = .07$)	4.05*
NoExpr $R^2 = .14$.82***	1.90 ($p = .05$)	2.59 ($p = .80$)	2.51	3.10 ($p = .07$)	3.35**	3.98***
Sad $R^2 = .41$.65***	2.87***	1.2***	1.49***	5.88	3.70***	5.41 ($p = .12$)
Solemn $R^2 = .30$.89***	2.5***	1.41***	2.16***	5.37***	3.78	5.01***
Tender $R^2 = .31$	1.06***	2.51***	1.6***	1.63***	5.82***	3.97 ($p = .06$)	4.60

Notes: The significance of the differences between the ratings for the intended expression and all other emotion ratings was assessed by a mixed effects model for each of the seven intended emotional expressions (see description in text). The effect size of each model (i.e. effect of type of emotion rated) was computed as marginal R^2 values for mixed effect models using the approach suggested by Nakagawa and Schielzeth (2013). Degrees of freedom for all significance tests are 2775. Only p -values for non-significant effects are reported. Reported significance levels are Bonferroni corrected. Significance levels are coded as: *** $p < .001$, ** $p < .01$, * $p < .05$.

Table A2. Model-based significant differences and mean emotions (columns) by intended expression (rows) for piano for Melody C.

IntExpr	Angry	Fearful	Happy	NoExpr	Sad	Solemn	Tender
Angry $R^2 = .31$	5.93	1.16***	2.26***	2.13***	1.40***	2.10***	.67***
Fearful $R^2 = .42$.60***	3.47	.91***	1.75***	6.02***	3.54 ($p = .81$)	6.12***
Happy $R^2 = .12$	2.48***	1.78***	4.16	2.23***	1.47***	2.35***	1.32***
NoExpr $R^2 = .22$	1.07***	2.35 ($p = .18$)	.98***	2.80	4.91***	3.44 ($p = .05$)	3.83**
Sad $R^2 = .36$.64***	2.77***	.84***	2.21***	5.99	3.55***	5.29*
Solemn $R^2 = .20$	1.12***	2.31***	.83***	3.25 ($p = .46$)	4.85***	3.50	3.29 ($p = .54$)
Tender $R^2 = .28$.90***	2.91***	.84***	2.67***	5.72***	3.64*	4.52

Notes: The significance of the differences between the ratings for the intended expression and all other emotion ratings was assessed by a mixed effects model for each of the seven intended emotional expressions (see description in text). The effect size of each model (i.e. effect of type of emotion rated) was computed as marginal R^2 values for mixed effect models using the approach suggested by Nakagawa and Schielzeth (2013). Degrees of freedom for all significance tests are 2775. Only p -values for non-significant effects are reported. Reported significance levels are Bonferroni corrected. Significance levels are coded as: *** $p < .001$, ** $p < .01$, * $p < .05$.

Table A3. Model-based significant differences and mean emotions (columns) by intended expression (rows) for violin for Melody C.

IntExpr	Angry	Fearful	Happy	NoExpr	Sad	Solemn	Tender
Angry $R^2 = .26$	5.67	2.34***	2.42***	.92***	1.74***	2.56***	1.23***
Fearful $R^2 = .05$	1.76***	3.02	2.31*	1.88**	2.93 ($p = .80$)	2.45 ($p = .11$)	3.43 ($p = .23$)
Happy $R^2 = .10$	1.79***	2.34***	3.95	1.42***	2.32***	3.10**	3.26*
NoExpr $R^2 = .06$	1.52***	1.83*	2.59 ($p = .80$)	2.68	2.87 ($p = .56$)	2.67 ($p = .98$)	3.67**
Sad $R^2 = .53$.74***	2.83***	.97***	.86***	7.18	4.83***	5.96***
Solemn $R^2 = .08$	2.04***	2.54**	2.45**	1.50***	3.36 ($p = .74$)	3.47	3.58 ($p = .72$)
Tender $R^2 = .36$.89***	3.23***	1.71***	1.50***	5.10**	3.64***	6.02

Notes: The significance of the differences between the ratings for the intended expression and all other emotion ratings was assessed by a mixed effects model for each of the seven intended emotional expressions (see description in text). The effect size of each model (i.e. effect of type of emotion rated) was computed as marginal R^2 values for mixed effect models using the approach suggested by Nakagawa and Schielzeth (2013). Degrees of freedom for all significance tests are 2775. Only p -values for non-significant effects are reported. Reported significance levels are Bonferroni corrected. Significance levels are coded as: *** $p < .001$, ** $p < .01$, * $p < .05$.

Table A4. Model-based significant differences and mean emotions (columns) by intended expression (rows) for voice for Melody C.

IntExpr	Angry	Fearful	Happy	NoExpr	Sad	Solemn	Tender
Angry $R^2 = .40$	6.74	1.66***	1.07***	2.32***	2.09***	1.51***	.55***
Fearful $R^2 = .47$.62***	4.47	.68***	1.58***	6.69***	3.63**	5.82***
Happy $R^2 = .18$	1.83***	1.11***	4.73	2.31***	1.37***	2.75***	1.75***
NoExpr $R^2 = .11$	2.70***	1.23***	1.81***	4.17	2.27***	2.13***	1.49***
Sad $R^2 = .50$.53***	2.9***	1.06***	1.35***	6.96	4.53***	6.03**
Solemn $R^2 = .29$	1.61***	2.22***	1.08***	1.5***	5.30 ($p = .10$)	4.77	3.34***
Tender $R^2 = .46$.36***	2.54***	2.17***	1.25***	5.00***	4.46***	7.26

Notes: The significance of the differences between the ratings for the intended expression and all other emotion ratings was assessed by a mixed effects model for each of the seven intended emotional expressions (see description in text). The effect size of each model (i.e. effect of type of emotion rated) was computed as marginal R^2 values for mixed effect models using the approach suggested by Nakagawa and Schielzeth (2013). Degrees of freedom for all significance tests are 2775. Only p -values for non-significant effects are reported. Reported significance levels are Bonferroni corrected. Significance levels are coded as: *** $p < .001$, ** $p < .01$, * $p < .05$.

Appendix 3. Table A5**Table A5.** ANOVA table (Wald χ^2 -tests, type III) of logistic mixed effects model of emotion decoding ability including experimental as well as individual differences factors.

	χ^2	<i>df</i>	<i>p</i>
(Intercept)	112.706	1	<.001 ***
Melody	3.791	2	.1502
Intended Expression	402.999	6	<.001 ***
Instrument	254.820	3	<.001 ***
Musical Training	30.263	1	<.001 ***

Significance level codes: ****p* < .001, ***p* < .01, **p* < .05.**Appendix 4. List of acoustical features****Table A6.** List of acoustical features.

Acoustical feature	Description
duration	Duration (s) (indicator of tempo)
brightness_mean	Mean of spectral brightness
brightness_std	Standard deviation of spectral brightness
roughness_mean	Mean of roughness
flatness_mean	Mean of spectral flatness
flatness_std	Standard deviation of spectral flatness
spec_entropy_mean	Mean of spectral entropy
spec_entropy_std	Standard deviation of spectral entropy
mode_mean	Mean of mode (0 = minor, 1 = major)
mode_std	Standard deviation of mode
pulse_clarity_mean	Mean of pulse clarity
pitch_mean	Mean of pitch (f0 extraction)
pitch_std	Standard deviation of pitch (f0 extraction)
kurtosis_mean	Mean of spectral kurtosis
kurtosis_std	Standard deviation of spectral kurtosis
rolloff85_mean	Mean of 85% spectral roll-off
rolloff85_std	Standard deviation of 85% spectral roll-off
rms_mean	Mean of root mean square amplitude (intensity)
skewness_std	Standard deviation of spectral skewness
zerocross_mean	Mean of zero-crossing rate
zerocross_std	Standard deviation of zero-crossing rate

Notes: Features were extracted with the MIRToolbox 1.6 for MATLAB. All features were calculated over 50 ms windows with 50% overlap, except duration (no window), mode (1 s windows with 50% overlap), and pulse clarity (5 s windows with 10% overlap).