A Brilliant, but Problematic Study: What do the Results of Frank (et al.) signify?

JOSHUA ALBRECHT [1]
The University of Iowa

ABSTRACT: Frank (et al.)'s article, "Exploring the variability of musical-emotion expression over historical time" is a thoughtful, creative attempt to tackle the thorny problem of reconstructing the historical cognition and perception of music. The article is a pleasure to read, owing in large part to the brilliant design of using musical excerpts composed by Johann David Heinichen to explicitly express particular emotional states. These descriptions serve as a ground truth for a modern empirical study. The study produced negative results, which are always tricky to interpret. This review explores some of the methodological choices of the study that may have contributed to the negative results. The review then questions the interpretations of the negative results suggested by the authors in light of some of the problematic methodological decisions of the study, especially the claims that the results suggest that the perception of musical affect has changed over the past 200 years. Finally, the review proposes what conclusions are appropriate to draw from the study in light of its design and framing.

Submitted 2024 February 12; accepted 2024 February 19.
Published 2024 June 7; https://doi.org/10.18061/emr.v18i2.9771

KEYWORDS: *historical music cognition, affect, methodology*

SOME types of studies immediately strike me for the creativity of the question they ask or the cleverness of their approach. "Exploring the variability of musical-emotional expression over historical time" succeeds on both levels. This article asks the provocative question of whether there has been any change in the way listeners decode musical affect over a period of hundreds of years from the Baroque era until today. The most direct way to test this change would be to bring in both eighteenth-century and modern *participants*, but of course, this is impossible. Instead, the authors ingeniously sample eighteenth-century and modern *excerpts*. Using these particular Baroque excerpts written by Johann David Heinichen is particularly clever because of the detailed descriptions the composer provided about what he intended the excerpts' emotional expression to be. These verbose descriptions serve as a sort of ground truth about the excerpts' intended emotional expression, even though there are necessary translation gaps between German and English and between sometimes long prose descriptions and the precise measures used in this study. Specifically, the authors map key affect words from the descriptions onto target quadrants on a 2D valence/arousal circumplex model. By asking modern participants to rate these excerpts on valence and arousal scales, the authors attempt to test whether modern listeners decode musical affect in ways that align with Heinichen's intentions.

I commend the authors of this study for the inventive idea of measuring modern listener perceptions of expressed emotion against Heinichen's descriptions. Finding these excerpts, matching them with their prose descriptions, and converting these descriptions into operationalized predictions for a modern empirical study is an original and thoughtful strategy for trying to solve the classic problem of understanding historical musical cognition, and I wanted it to work.

However, I believe that there are some fundamental issues with the methodology and theoretical framing of this study that undermine its laudable aspirations. There are three main issues that I'd like to discuss below: the problematic measurement tool used, the oversimplification of musical cue/affect correlation, and the mis-framing of the evolutionary perspective of musical affect.

The study produced negative results, at least regarding predictions of modern participants' ratings of Heinichen's excerpts. There are all sorts of reasons why these negative results may have occurred, including the authors' bold claims that "[t]hese results may be indicative of systematic changes in listeners' perception of musical emotions over time" (p. 123). However, given some of the methodological issues in

the study, it is my opinion that the results are at least as likely to be indicative of other less exciting conclusions (discussed below). In light of the article's strongly worded conclusions, it bears emphasizing that negative results are a lack of evidence for the predicted effect and not evidence for the opposite of the predicted effect, as suggested by the authors. The distinction is a subtle but important one. Even though the results are quite interesting, and I'm glad that the study and its results are being published, I'm not sure that they reveal much about the primary hypothesis—despite the authors' claims to the contrary.

MEASUREMENT TOOL ISSUES

In the abstract and throughout the article, the authors describe the "significant mismatch between original descriptions and listener ratings" (p. 117). However, I'm not confident that the results are evidence of a significant mismatch, owing largely to the measurement tool used. There are three primary issues with the measurement tool used.

First, as the authors acknowledge, there are significant challenges in boiling down Heinichen's verbose prose descriptions to one core affect. Translating from German to English is not always without loss of some meaning, but the bigger issue is reducing the rich prose Heinichen provides to one single affect. Consider his description of the E4 excerpt, for example: "Should one wish to try special expressions, the words *faville*, *pupille*, *l'ardore*, *lo squardo* give our imagination much opportunity for pleasant and almost playful inventions. For example, one could represent the burning fire of **love** in the following invention" (Appendix A). By focusing on just the notion of "love," the authors concluded that this would be a low arousal, positive valence excerpt. Though "love" may often be low arousal, in this case, "the burning fire of love," would suggest a passionate, tempestuous form of love, hardly low arousal. Based on the totality of the quote, choosing low arousal is problematic. Indeed, the results (Figure 2) reveal positive valence and high arousal. While the authors predicted low arousal and consider this a predictive failure, given the provided description it would be just as easy to conclude that modern listeners rated this excerpt *consistent* with Heinichen's description.

This leads to the second main problem with the measurement tool used. The authors admit in the article that there is a lot of subjectivity in getting from the prose descriptions to quadrants on the valence/arousal plane, something Heinichen would have likely found peculiar. There is no fundamental issue with measuring arousal and valence, a standard response measure that is certainly helpful in comparing the results with other studies. But if the authors wanted to know whether modern listeners still heard the affects that eighteenth-century listeners would have presumably heard (or at least the one eighteenth-century listener Heinichen), they should have just asked. They could have directly measured participant responses on a Likert scale for the terms that came out of Heinichen's prose. For example, participants could have been asked the degree to which excerpts expressed the affects "playful," "furious," or "in love" (or "tender" or some other appropriate adjective), or even "burning fire" or "passionate." If these measures did not align with Heinichen's descriptions, then there would be much stronger evidence for the authors to conclude that listeners were not judging the excerpts in the same way.

Finally, as an abstract representation, valence/arousal can be imprecise when measuring specific affects. For example, in Figure 1 (p. 120), one might imagine that "anger," "fear," and "disgust," while all properly being Q2 terms, might still be experienced as categorically different. Alternative axes or three dimensions (e.g. adding dominance) are sometimes used to get around these issues. Even how participant responses were measured is problematic; happy/sad and energetic/sleepy as opposite poles leads to further precision issues. While strongly negatively correlated, sad and happy appear not to be direct opposites, perhaps owing to the implied arousal in the terms (happy being higher arousal than sad). In a study in which participants rated happy/joyful and sad/depressed/tragic independently in Romantic piano music, I found ratings of these ostensibly opposite affects to only be negatively correlated at -.635 (Albrecht 2012: 104). While the terms of that study are more complex than simply happy and sad, in a follow-up study with the simpler terms happy and sad, I found an even weaker correlation at -.596 (Albrecht 2016, see Figure 1). Even sleepy and energetic may not be opposites owing to the implied valence embedded in those terms. More careful Likert scale usage would measure the degree to which a single affect is present with both poles labeled with the same term. Altogether, these issues with the measurement tool raise questions about what the results signify.

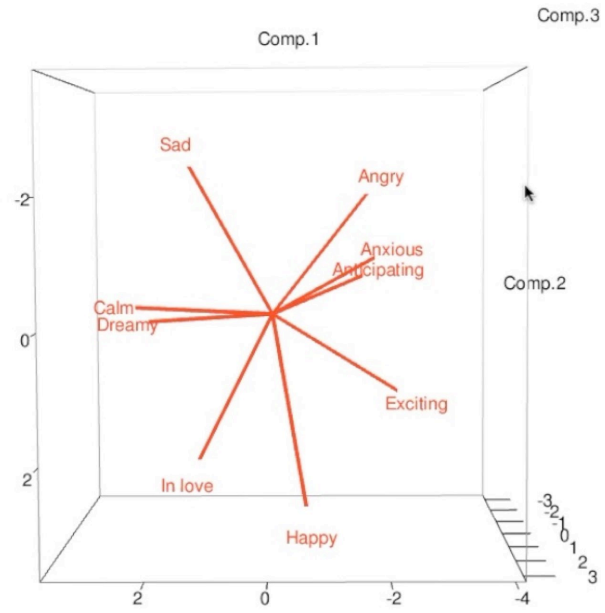

Figure 1. The results of a study asking participants to rate Romantic piano music using discrete affect categories independently (Albrecht, 2016). Principal components analysis reveals that while happy and sad ratings are highly negatively correlated, these terms are not diametrically opposed, with a negative correlation of only -.596.

OVERSIMPLIFICATION OF MUSICAL CUE AND AFFECT CORRELATION

What accounts for the discrepancy between predicted participant ratings and actual ratings? Beyond methodological issues with the measurement tool used, the authors explore the possibility in the discussion that Heinichen's musical examples may have shifted in meaning over the centuries. Though one may reasonably question whether the results "suggest that modern-day listeners interpret Heinichen's examples differently to the composer's intentions" (p. 123), it is nonetheless worthwhile to consider what may have changed from the eighteenth century until today. By exploring "[s]hifting patterns of relationships between musical-structural emotion cues…and changes in average cue levels," the authors attempt the difficult but important task to "differentiate between the universal and the culturally variable in musical emotion" (p. 123). They then point out several ways in which prior research has investigated changes in musical cue usage over historical time, focusing specifically on the role of mode, the rate of event presentation (REP, or rhythmic onset speed), mean pitch height, and mean melodic interval size.

While it is helpful for the authors to explore the relationship between musical cue usage and expression, their analysis of that relationship tends to be a bit oversimplified. The primary issue is that this article examines the relationship of musical cues in isolation without considering how they might be influenced by other variables. Consider the case of pitch height. The authors observe that prior research has found positive correlations between pitch height and arousal but negative correlations between pitch height and valence and that consequently the picture is muddied when a cue is associated with both valence and arousal (p. 124). How then to understand negatively valenced, high arousal emotions like furious? The answer, of course, is that the usage of one musical cue is mediated by other musical cues. Albrecht (2016) found that excerpts expressing angry affects exhibited both lower lowest pitch, but also higher highest pitches (see the article's Tables 7 and 9). In other words, overall tessitura expanded, consistent with increasing arousal, as opposed to a constricted tessitura found in excerpts expressing calm (i.e. a low arousal, neutral/positive valence affect). Lower lowest pitches were also found in excerpts with high arousal, negative valence affect in general, consistent with acoustic attributes of threat displays in speech. In contrast, higher highest pitches depended on the specific affects expressed, resulting in a nuanced interaction effect of pitch height on both arousal and valence. Dynamic level and mode complicated the issue further, with the specific affect expressed by lower pitch height being mediated by the mode and loudness of the excerpt.

Although there do seem to be main effects of specific musical cues on valence or arousal averaged across all other musical cues, in reality the story of musical affect is more complicated than main effects. Affect is an emergent property that arises out of the interaction of many different musical cues in conjunction. So, the authors speciously oversimplify the matter when they conclude that "E4 (love) has the highest mean pitch of the three examples, which does not match expectations based on Heinichen's descriptions" (p. 124). Though pitch height and interval size may in general be positively correlated with arousal, this does not necessarily mean that we should expect low pitch height in this particular example. As pointed out above, love is already complicated enough to not easily be pinned to a specific quadrant on the circumplex model and in this particular case may indeed be high arousal. But even then, the relationship between high pitches and arousal is itself complicated enough to not have a direct 1:1 correlation with pitch height generally, but is mediated by interactions with other musical cues. In summary, finding high pitches in one "love" excerpt is not sufficient evidence to conclude that musically expressed love has changed over the intervening years.

To be fair, the authors offer a disclaimer at the end of this paragraph: "these expectations are made based on the VA mappings of the emotion keywords for each excerpt, rather than the emotion terms themselves." Nevertheless, the oversimplification of cue/affect correlation characterizes the discussion throughout. As another example, consider the discussion of the relationship between onset speed (called REP) and the excerpt expressing playful (E1). The authors first conflate onset speed (REP) with tempo and then identify it as being always positively correlated with valence and arousal: "Tempo, alongside mode, is another cue that shows remarkable consistency…associated positively with both valence and arousal…REP levels in the excerpts do not fit expectations…" (p. 124). In the first case, tempo and onset speed are not the same thing, with shorter notated durations often more common in slower tempos and vice versa. In the specific case of playful, it is unclear where the expectation that playful should have a high REP comes from. It seems that it's because the authors assigned playful to Q1 with high arousal. But playful *may* have high arousal. Although there isn't a large body of research examining musical playfulness specifically, in its musical expression it is reasonable to assume that playful would have more *variability* of rhythmic duration rather than simply high onset speed, as reflected in a measure like nPVI (see e.g., Patel & Daniele, 2003). I often hear music with grace notes, dotted or doubly-dotted rhythms, or triplets as playful, especially when these rhythms are in contrast with other durations (e.g. the "Scotch snap," see Temperley & Temperley 2011). To be sure, this would be a hypothesis needing formal testing, but the playful musical example does reflect rhythmic variability. The simple assumptions based on quadrant-placing single terms reveal the oversimplifications the authors make with these complex original descriptions.

EVOLUTIONARY RELATIONSHIP BETWEEN MODE AND AFFECT

Probably the strongest example of oversimplifying cue usage is found in the article's treatment of mode. The authors helpfully contextualize the affective meaning of mode within its historical context, pointing out on the first few pages that mode as a binary (major/minor) grew out of a larger collection of church modes. Therefore, mode as a binary reflecting a loose mapping of valence as a binary must not be a timeless phenomenon, but at least limited to the historical development of mode as a binary. This, without needing to collect any participant response data, represents a historical argument for their primary thesis, revealing at least one instance of the affective meaning of musical cues shifting over time owing (at least) to the *use* of one musical cue (mode) shifting over time.

Still, the authors seem to suggest the argument that theorists, especially evolutionary theorists, consider affective meanings of mode to somehow be inherent to mode or humanity due to evolutionary forces. For example, they claim that "[s]ome [evolutionary] researchers have nonetheless suggested candidate rationalizations for the emotional impact of mode, for example noting that melodies in the minor mode tend to consist of smaller intervals than those in the major mode, with these intervals potentially eliciting negative valence" (p. 118). Later, they conclude that "[s]uch theories would demand that the major/minor binary (where it exists) be universally valenced, implying a historically static interpretation of mode" (p. 123). These assumptions that the minor mode should always be negatively valenced leads them to speciously conclude later that their results suggest "that emotional connotations of modes have changed over time" (p. 123).

However, the evolutionary rationalization for the minor mode expressing sadness is more complex than simply suggesting that small intervals are somehow inherently negatively valenced, as evidenced through much of David Huron's research. Huron (2015) argues that many of the musical cues traditionally associated with sad affect are a reflection of low physiological arousal rather than simply valence. Low physiological arousal is associated with low acetylcholine and low epinephrine, which results in lower sub-

glottal air pressure and less tense vocal folds with looser muscle tension in lips and tongue, in turn resulting in smaller pitch movements (Huron & Davis, 2012), lower pitch (Huron 2008), and quieter dynamics (Turner & Huron, 2008).

Small pitch movement, a hallmark of minor-mode music (Huron & Davis, 2012), is therefore associated with low physiological arousal in general, which could be either negatively valenced or positively valenced, such as in a state of relaxation or peacefulness. So the predominance of small intervals in some work is not necessarily indicative of negative valence by itself. Instead, small intervals merely suggest lower overall arousal, and what valence the music takes on would be influenced by other musical factors. If lower pitch height alone were indicative of sadness, then male speakers should tend to all sound depressed compared with female speakers. However, prosody research shows that listeners "normalize" speaker pitch, so sadness is perceived only when the pitch is lower than expected or "lower than normal" for a given voice (Huron, Yim, & Chordia, 2010).

So how does a listener "normalize" a musical context? Today, roughly 75% of Western Classical music is in the major mode (Albrecht & Huron, 2012). For modern ears, the minor mode is experienced as an exception to the primary major-mode norm. As Temperley and Tan (2013) have shown, as the number of major scale pitches are lowered, modern listeners hear the resulting mode as increasingly sad, and the effect is not limited to ♭3 and ♭6. In short, the perceptual research suggests that there is nothing inherently sad about the minor mode. Rather, the operative principle is that "lower than expected" or "lower than normal" tends to be perceived as sadder. Therefore, whether a given mode in some culture is heard as "sad" depends on how that mode differs from the culturally normative or most commonly experienced scale in that culture. One method of estimating the affective meaning and valence a particular mode might have in a historical period, then, would be to establish what is statistically "normative" music. The hypothesis would then be that music or modes lower than that normal would be heard as sad.

A further wrinkle is that prosodic research also links lower-than-normal pitch in speech with aggression—but only when the sound is intense rather than quiet. This same effect is evident with the minor mode. Hevner (1935) noted that loud minor-mode music is perceived as serious, passionate, or aggressive, rather than sad. A good example of this would be the opening minor chord at the beginning of Beethoven's Pathetique sonata: "serious" or even "aggressive" are certainly better descriptors than "sad." Specifically, Horn and Huron (2015) explicitly take as the premise of their study that affect is an emergent property of the interaction of several musical factors. They argue that minor is not inherently sad, but only when it is presented at a low dynamic, at a slow tempo, and with more legato articulations. Recalling Hevner (1935), they argue that fast, loud, staccato minor is not sad at all and may not even be negatively valenced, but is instead "passionate." The change they discovered was the greatly increased proportion of passionate minor mode music into the 19th century.

So, when this article claims that "[c]hanges in cue use over large corpora may reflect the changing prominence of different emotion portrayals (as concluded by Horn & Huron, 2015), but could additionally hint towards changes in the emotional impact of the cues" (p. 123), the authors mischaracterize the results of that study. On the contrary, Horn and Huron's study is predicated upon stable associations between musical-structural emotion cues as an interaction between at least four musical cues. Under the right circumstances, there are plenty of instances of positively-valenced minor mode music and negatively-valenced major mode music. The results of this study, then, do not provide enough evidence to conclude that musical affect has significantly changed in the intervening years, especially given the small number of musical examples surveyed and the simple correlation examined between mode and affect without taking some of the nuances of evolutionary arguments of musical affect into account.

CONCLUSION

This article presents a great study, one that is thoughtful, inventive, and creatively seeks to answer the question of how the perception of musical affect has changed over historical time. The premise is intriguing, the strategy of answering its posed questions is commendable, and the results are interesting. Looked at with the right lens, these results can inform a great deal about the perception of musical affect, and even historical musical affect perception. There is certainly a lot to like in a study like this, and I am glad that the article is now part of the scholarly conversation about historical music cognition.

Given all its positives, the challenge in understanding what this article offers the reader has been sufficiently outlining what we can learn from the results, given the problematic nature of some of the methodological decisions made and what meaning the authors extract from the results. The essential issue is

that authors have gone too far in suggesting their results are evidence of a monumental shift in musical perception. These are results on perceptions of a few excerpts from a single composer using distant representations (arousal/valence quadrants) of complex prose descriptions provided only by the composer's own hand. By suggesting that, "[s]hould the results of this study be generalizable, such "appropriateness" [that a performer's interpretation is connected to a performer's intention] may become more difficult to assert," (p. 123) the authors go too far in claiming a potential paradigm shift in interpreting historical musical affect based on their findings. While the negative results could signal that there has been a significant shift in affective meaning of Baroque music from then until today, other less interesting implications seem more likely.

Given the methodological limitations, it seems just as likely that the negative results could be because the VA Likert task maps poorly onto Heinichen's prosaic descriptions. Or they could suggest that equal temperament ruins the affective connotations of the music (as Heinichen himself may have argued). Or it could suggest that a dynamic-less performance loses something. It could be that Heinichen was just an ineffective composer, at least in the emotional domain. It could be that positive and high arousal ratings may simply be given as defaults to what is perceived as default music such as generic Baroque excerpts, especially when compared against modern excerpts.

At the most basic level, all the authors can conclude is that modern listeners do not rate Heinichen excerpts in the 2D valence/arousal quadrants that the authors expected based on their interpretations of his qualitative descriptions. If the results are examined in this light, I think there is much to be learned from them. Future research could creatively tease apart some of these possibilities to get closer to understanding why the results of this study were as they are. There are undoubtedly reasons these particular excerpts were evaluated in these particular ways, both by modern listeners and by Heinichen himself, and research that leads to such intriguing questions as this article has offered promises for a fascinating future of historical music perception.

ACKNOWLEDGEMENTS

This article was copyedited by Eve Merlini and layout edited by Jonathan Tang.

NOTES

[1] Correspondence can be addressed to: Dr. Joshua Albrecht, The University of Iowa, 5415 Voxman Hall, 93 E. Burlington St., Iowa City, IA 52246, joshua-albrecht@uiowa.edu.

REFERENCES

Albrecht, J. (2016). Learning the language of affect: A new model predicts perceived affect across Beethoven's piano sonatas using music-theoretic parameters. *Proceedings of the ICMPC-SMPC 2016 Joint Conference* (pp. 79-85). San Francisco, CA.

Albrecht, J. (2012). A model of perceived musical affect accurately predicts self-reported affect ratings. *Proceedings of the 12th International Conference of Music Perception and Cognition* (pp. 35-43). Thessaloniki, Greece: ICMPC.

Albrecht, J. & Huron, D. (2012). A statistical approach to tracing the historical development of major and minor pitch distributions, 1400-1750. *Music Perception: An Interdisciplinary Journal 31*(3), 223-243. https://doi.org/10.1525/mp.2014.31.3.223

Hevner, K. (1935). The affective character of the major and minor modes in music. *American Journal of Psychology, 47*, 103-118. https://doi.org/10.2307/1416710

Horn, K & Huron, D. (2015). On the changing use of the major and minor modes 1750-1900. *Music Theory Online, 21*(1). https://doi.org/10.30535/mto.21.1.4

Huron, D. (2008). A comparison of average pitch height and interval size in major- and minor-key themes: Evidence consistent with affect-related pitch prosody. *Empirical Musicology Review*, *3*(2), 59-63. https://doi.org/10.18061/1811/31940

Huron, D. (2015). Cues and signals: An ethological approach to music-related emotion. *Signata 6*(6), 331-351. https://doi.org/10.4000/signata.1115

Huron, D., & Davis, D. (2012). The harmonic minor scale provides an optimum way of reducing average melodic interval size, consistent with sad affect cues. *Empirical Musicology Review, 7*(3-4), 103-117. https://doi.org/10.18061/emr.v7i3-4.3732

Patel, A. & Daniele, J. (2003). An empirical comparison of rhythm in language and music. *Cognition, 87*, (pp. B35-B45). https://doi.org/10.1016/S0010-0277(02)00187-7

Temperley, D., and Tan, D. (2013). Emotional connotations of diatonic modes. *Music Perception: An Interdisciplinary Journal, 30*(3), 237-257. https://doi.org/10.1525/mp.2012.30.3.237

Temperley, N. & Temperley, D. (2011). Music-Language Correlations and the "Scotch Snap." *Music Perception: An Interdisciplinary Journal, 29*(1), 51-63. https://doi.org/10.1525/mp.2011.29.1.51

Turner, & Huron, D. (2008). A comparison of dynamics in major- and minor-key works. *Empirical Musicology Review, 3*(2), 64-68. https://doi.org/10.18061/1811/31941