A Call for Hypothesis-Driven, Multi-Level Analysis in Research on Emotional Word Painting in Music: Commentary on Sun & Cuthbert (2018)

NIELS CHR. HANSEN[1]
MARCS Institute for Brain, Behaviour, and Development, Western Sydney University

ABSTRACT: This commentary discusses Sun and Cuthbert's (2018) exploratory analysis of emotional word painting in a corpus of English-language popular and folk songs. The authors are complimented for their application of computational tools to an impressively large sample of a somewhat understudied musical genre, and for their detailed level of analysis mapping musical features to the semantic content of individual words. This work, however, suffers from a lack of *a priori* predictions which causes multiple comparison issues leading to a dramatic reduction in statistical power. The selection of musical features and analytical strategies also seems arbitrary at times due to the absence of motivating hypotheses. It is argued that the ethological literature on affective vocal communication in animals might offer an avenue for future hypothesis-driven research on this topic.

Submitted 2018 September 15; accepted 2018 October 1.

KEYWORDS: *emotion, affect, sentiment, ethology, lead sheet, lyrics, corpus studies, NRC EmoLex, Wikifonia, music21*

> *Much of the work of the film composer is a kind of musical alchemy, pouring rarified ingredients (and more than a drop of our own blood) into a bubbling cocktail of pitches, patterns, modes and memories. We can nudge around neutralities or at extremes, overtly manipulate raw emotions, all the while not really knowing how or why this might make a person cry […] Indeed as composers, we are not painting a full and complete picture for our listeners of what to feel, we are really just opening a window and pointing in a certain direction. (Douek, 2013)*

MANY adept music listeners will resonate with these thought-provoking words written by the English-born, American-based film composer Joel Douek when asked to describe what the creative process of composing music really entails. Some aspects of musical composition can indeed be viewed as a process of setting emotions to music or painting one's emotions in the colors of music. This may, in turn, lead to popular opinions that "music is what feelings sound like" (Warren, 2014), which occasionally infiltrates scholarly discourse (e.g., Stouten, Glissen, Camps, & Tuteleers, 2012). Despite an abundant literature educating aspiring students in the craft of musical composition, a grain of truth remains in Douek's provocative claim that the scientific understanding of how human emotions are effectively set to music is only marginally more developed than that of medieval witchcraft.

The recent paper "Emotion painting: lyric, affect, and musical relationships in a large lead-sheet corpus" by Sophia S. Sun and Michael Scott Cuthbert (2018) takes an important step towards illuminating the issue of emotional word painting in music. The most interesting conclusions drawn from this exploratory corpus study include that negatively valenced lyrical emotions have a stronger tendency than positively valenced lyrical emotions to occur on metrically weak positions, employ shorter note durations, and show more ambiguous key clarity. Positively valenced lyrical emotions may, on the other hand, more often coincide with higher-pitched notes and be approached by larger melodic intervals. Words associated with surprise may, furthermore, coincide with more dissonant pitches and minor chords.

While Sun and Cuthbert's (2018) correlational approach naturally calls for empirical validation in listening experiments with systematic manipulation of candidate musical features, the authors should be

complimented for their capitalizing on computational tools such as the *music21* toolkit for Python. This enables them to analyze an impressively large sample of 1,895 folk and popular songs with English-language text providing 237,428 musical notes, 69,498 chord symbols, and 205,724 candidate words from the accompanying song lyrics.

Comparable large-scale studies have previously related musical features to the proportional occurrence of words associated with affective dimensions such as "anxiety", "anger", and "sadness" (DeWall, Pond, Campbell, & Wenge, 2011; Léveillé Gauvin, 2018). However, these earlier analyses have almost exclusively been done on the level of entire songs rather than individual words.

Previous studies have typically used different versions of the *Linguistic Inquiry and Word Count* (LIWC) program (Pennebaker, Boyd, Jordan, & Blackburn, 2015) which relies on secret, proprietary word lists. Sun and Cuthbert (2018), on the other hand, made use of the *NRC Word-Emotion Association Lexicon* (EmoLex) (Mohammad & Turney, 2013), which provides a freely available online vocabulary of affectively annotated words in a multitude of languages.[2] The same admirable spirit of resource sharing is evident from the author's decision to provide open access to their code and a complete list of song titles on Github.[3] Unfortunately, the musical corpus itself is not available.

WORD-BASED LEVEL OF ANALYSIS

In addition to the use of a large musical sample, the primary novel contribution of this paper pertains to its detailed level of analysis. Specifically, musical features are related directly to the affective content of co-occurring individual words rather than summarizing semantics across entire songs. To my knowledge, Paul and Huron's (2010) small-scale study of breaking voice and grief-related content in country music lyrics constitutes the only prior example of empirical research conducting detailed word-level analysis in this way. Their finding that the association of breaking voice and grief was only present at the level of individual words, but not at the level of individual lines, stanzas, or entire songs suggests that Sun and Cuthbert's (2018) decision to focus on the word level was indeed justified.

Yet, while word-based analysis allows the authors to detect more fine-grained affective communication in music, the possible pitfall remains that this detailed level of analysis is too local and that, consequently, other pertinent features may have been ignored. This may potentially underlie the surprising lack of emotional valence-related effects on pitch height, which could easily spill over into neighboring events or span longer melodic phrases while still constituting true cases of emotional word painting.

An example of the possible inadequacy of word-level analysis appears in the authors' own account of potential madrigalisms in Elvis Presley's performance of the hook line from George Weiss' 'Can't Help Falling in Love'. Here, it is concluded that the melody violates semantic expectations by ascending stepwise on the word "falling". However, the fact that pitch descent does indeed occur immediately afterwards on the word "love" suggests that a larger unit of analysis may have been more relevant.

Alternatively, "falling" could be interpreted as a reference to romantically induced vertigo or imbalance. This would be aptly illustrated by the use of triplet rhythms setting it aside from the immediate musical context. In recently published work, David Huron and I provide an ecological theory explaining the qualia of rotation and musical rhythm (Hansen & Huron, 2019). We argue that triplet rhythms are especially effective in conveying a sense of rotation or spinning due to their convergence with loudness trajectories generated by truly rotating, sound-emitting objects in the environment. Empirical data collected from listening experiments seem largely consistent with this theory.

Because it remains unresolved what the optimal level of analysis is for understanding emotional word painting or madrigalisms in Western pop and folk musics, converging empirical approaches based on theory-driven predictions should be encouraged in future work. This brings us to the pertinent issue of hypothesis-driven and exploratory research.

HYPOTHESIS-DRIVEN VS. EXPLORATORY RESEARCH

In my view, the primary shortcoming of Sun and Cuthbert's (2018) paper relates to its highly exploratory nature with a complete lack of *a priori* hypotheses. Instead of acknowledging this, the authors seem to occupy a somewhat arbitrary middle ground between a properly Bonferroni-corrected exploratory study and a conclusive research study testing a limited number of predefined hypotheses. All in all, it is not clear which of these two camps this work fits into. Although some references to "hypotheses" do appear in the results section (e.g., in the "major/minor chord context" and "mode" sub-sections), the fact that they are

never explicitly stated prevents research design and statistical test procedures from being optimized accordingly. Consequently, while this research may raise a range of interesting questions to test in future studies, regrettably, it fails to provide any definitive answers.

A number of unfortunate consequences tend to arise from the absence of predefined hypotheses. First of all, the work suffers severely from multiple comparison issues, leading to dramatic reductions in statistical power. In total, 85 statistical tests are conducted for each of the 12 examined correlators (beat strength, note length, relative pitch height, pitch interval, chord consonance, chord membership, chord modality, key clarity, overall mode, moving-window mode, pitch class, and word length). If one were to Bonferroni-correct using the conventional 95% confidence level, the alpha level would need to be reduced to .05/1020 = .000049. This would diminish statistical power considerably and few results would reach the threshold for statistical significance. This is especially the case for those results reported in Figs. 10, 11, 12, 13, and 14.

On a practical level, strategies for at least partly mitigating these multiple comparison issues do exist. For example, prior to conducting the study, the authors could have committed to a specific vocabulary (rather than using four different ones) as well as to a single operationalization of musical features like consonance and mode. Similarly, it seems debatable how meaningful the many pairwise comparisons across all affective dimensions really are. Pilot analyses using a different sample could relatively easily have provided the necessary validation to justify more selective methodological decisions. This would have increased statistical power considerably with limited reduction in explanatory power.

Importantly, the overall lack of statistical power somewhat compromises the interpretation of null results regarded by the authors as "contradicting previous research" (p. 327). This is the case for pitch height, consonance, and mode, which were expected to be more clearly associated with emotional valence. Moreover, some findings with relatively low *p*-values cannot be trusted. The purported association of joyous and positive lyrics with minor mode, for example, does not by any means pass correction for multiple comparisons. Comparably sized *p*-values occur in the pitch-class analysis referred to by the authors as a "meaningless metric" (p. 343) derived to check the validity of their analyses.

Sun and Cuthbert's (2018) approach generally results in a disproportionate focus on statistical significance over practical significance. For example, the authors argue that "the difference in average pitch for notes setting words connected to 'anger' compared to those set to emotional-neutral words might be less than a semitone, but if such difference is consistent throughout the thousands of examples, it could still be highly significant" (p. 328). While this may be true in a statistical sense of the word "significance", it has little relevance to actual music perception if individual participants in a single listening context are unable to detect such subtle differences in overall pitch averaged across thousands of songs. The same could be said about the finding that "sad" words are sung less than half a semitone higher, on average, than "angry" words. In these cases, local pitch differences, as quantified in terms of the pitch interval preceding the affectively pertinent note, may be a cue that is more realistically picked up by actual listeners. The absence of a difference in preceding interval between "sad" and "angry" words may thus be far more significant in a practical sense.

Finally, because they may have been chosen based on what *music21* had to offer rather than on predefined hypotheses, some of the musical features analyzed in this study seem potentially suboptimal. Although suggestions of other alternative features like pitch direction, rhythmic regularity, and melodic consonance are made in the discussion section, the crucial point is that these are not backed up by motivating hypotheses.

Rather than studying major and minor modes on a single continuum, the relationship between semantic content and "majorness" as well as "minorness" could, for example, be assessed independently, as we did in a recent corpus study on the emotional implications of solo instrumentation in Western orchestral music (Hansen & Huron, 2018). It is indeed possible that minor—being the more "unusual" key in much tonal music—could be more effectively used to mark emotional content than the run-of-the-mill major.

Moreover, the procedure of counting quarter notes and relating these absolute values to a distribution of notated durations across the entire corpus does not seem sufficiently sensitive to within-song variations within this particular musical parameter. Note length may indeed follow affective patterns whose scale is specific to individual songs. Tempo is also not controlled for. It is certainly not impossible that notes with the same nominal note value may convey dramatically different emotions when played in a fast versus a slow tempo. While pitch height was normalized for each song, this was curiously not done for pitch interval, note length, consonance, and the various key-related features. On the other hand, within-song

normalization would assume that individual songs span a range of affective dimensions rather than focusing on just one. This assumption would obviously not always be tenable either.

In summary, while advancement of any scientific field may benefit from a suitable combination of exploratory and hypothesis-driven research approaches, the former should primarily be reserved for subject matters where prior results provide limited grounds for formulating *a priori* hypotheses. As I will argue below, this is not entirely the case with regards to the topic of emotional word painting in music.

POSSIBLE SOLUTIONS

In criticizing the lack of hypotheses, it seems relevant to discuss where relevant hypotheses could have come from. While the authors mention Gabrielsson and Lindström's (2001) seminal article, they fail to derive *a priori* hypotheses from this review. As outlined by Hansen (2013), the rich research literature on emotions in music spans back to at least Hevner (1937) and Rigg (1937) with a multitude of other papers published during the intervening years before Gabrielsson and Lindström's overview. Although their 17-year-old article provides an important contribution to the field, it is also not fair to claim that it still represents "the current state of research on the perceived emotional content of individual musical elements" (p. 328).

The authors seemed puzzled that nominally "sad" and "negative" words are not preceded by descending pitch intervals in their corpus of popular and folk songs. However, while generally low pitch has been established as an acoustic cue for sad affect, narrow pitch range and lack of large interval leaps are equally common cues for sadness (Huron, 2008). If half of such small intervals ascend and the other half descend, the mean absolute interval preceding sad words would center around zero. In other words, although sad words are not more typically approached from above than from below, musical emotion painting may indeed be happening for these words if one considers the low variability of F0 found in sad speech and sad music.

The ethological literature on affective vocal communication in animals may offer the most interesting and relevant perspective on the issue of emotional word painting in music. Huron (2015), in particular, has discussed this possibility in great detail, purporting that many musical sounds emulate vocal displays that either inadvertently give away cues about the adaptive fitness of the observed individual or serve to provide affective signaling to the benefit of the signaling individual. According to Huron's (2015) 'Acoustic Ethological Model'—which expands upon Morton's (1977; 1994) prior work—quiet and high-pitched sounds are associated with appeasement and friendliness, loud and high-pitched sounds are associated with fear or alarm, quiet and low-pitched sounds are associated with sadness, relaxation, or sleepiness, whereas loud and low-pitched sounds are associated with aggression or seriousness. While empirical tests have established that some of these principles may apply to various musical corpora (e.g., Huron, Kinney, & Precoda, 2006), their applicability to Western pop and folk musics remains somewhat understudied. In addition to biologically motivated sources of hypothesis formulation, findings relating to historical changes in the use of musical parameters may also be of interest (e.g., Hansen, Sadakata, & Pearce, 2016).

The authors provide an interesting musical analysis of possible madrigalisms in Franz Schubert's lied 'Gretchen am Spinnrade' and songs by Elvis Presley. They note the apparent inconsistency that joyful words like *Herz* ("heart"), *Lächeln* ("smile"), *Zauber* ("magic"), and *Küss* ("kiss") in Gretchen's singing fall on high pitches whereas Elvis tends to sing words like "love" and "darling" on relatively low pitches. Ethological signaling theory may indeed have interesting things to say about these issues. According to Huron's (2015) 'Acoustic Ethological Model', for example, high pitch may be used to convey submissiveness whereas low pitch may be used to signal aggressiveness. Sound-size symbolism, in other words, predicts that raising your pitch makes you sound smaller (and less threatening) whereas lowering it will make you sound larger (and more powerful). Both Gretchen (by courtesy of Schubert) and Elvis may thus be adhering to ethological principles (and possibly Victorianesque sexual morality) when the former signals submissiveness towards a potential romantic partner whereas the latter signals capability of protecting the other person.

CONCLUSION

In conclusion, Sun and Cuthbert's (2018) study provides an important step forward in terms of adopting computational tools to the analysis of large musical corpora, by extending previous work to understudied

musical genres, and by focusing on a detailed unit of analysis. Its lack of predefined experimental predictions, however, renders this work somewhat inconclusive. Future research should strive to take advantage of the existing research literature on musical emotions by adhering to strictly hypothesis-driven, multi-level analysis with the purpose of properly uncloaking the wizardry of emotional word painting in music.

ACKNOWLEDGMENTS

This article was layout edited by Diana Kayser.

NOTES

[1] Correspondence can be addressed to: Niels Chr. Hansen, MARCS Institute, Western Sydney University, Locked Bag 1797, Penrith NSW 2751, Australia, Ni3lsChrHansen@gmail.com.

[2] See https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

[3] See https://github.com/cuthbertLab/emotionPainting.

REFERENCES

DeWall, C. N., Pond, R. S., Jr., Campbell, W. K., & Twenge, J. M. (2011). Tuning in to psychological change: Linguistic markers of psychological traits and emotions over time in popular US song lyrics. *Psychology of Aesthetics, Creativity, and the Arts*, 5(3), 200-207. https://doi.org/10.1037/a0023195

Douek, J. (2013). Music and emotion—a composer's perspective. *Frontiers in Systems Neuroscience*, *7*, 82. https://doi.org/10.3389/fnsys.2013.00082

Gabrielsson, A., & Lindström, E. (2001). The influence of musical structure on emotional expression. In P. N. Juslin and J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 223-248). Oxford, UK: Oxford University.

Hansen, N. C. (2013). Cognitive approaches to analysis of emotions in music listening. In: M. Zatkalik et al. (Eds.), *Histories and narratives of music analysis* (pp. 597-627). Cambridge, UK: Cambridge Scholars Publishing

Hansen, N. C., & Huron, D. (2018). The lone instrument: musical solos and sadness-related features. *Music Perception*, *35*(5), 540-560. https://doi.org/10.1525/mp.2018.35.5.540

Hansen, N. C., & Huron, D. (2019). Twirling triplets: the qualia of rotation and musical rhythm. *Music & Science*, 2. https://doi.org/10.1177/2059204318812243.

Hansen, N. C., Sadakata, M., & Pearce, M. (2016). Nonlinear changes in the rhythm of European art music: Quantitative support for historical musicology. *Music Perception*, *33*(4), 414-431. https://doi.org/10.1525/mp.2016.33.4.414

Hevner, K. (1937). The affective value of pitch and tempo in music. *The American Journal of Psychology*, *49*, 621-630. https://doi.org/10.2307/1416385

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). Affect induction through musical sounds: an ethological perspective. *Philosophical Transactions of the Royal Society of London B*, 370(1664), 20140098. https://doi.org/10.1098/rstb.2014.0098

Huron, D., Kinney, D., & Precoda, K. (2006). Influence of pitch height on the perception of submissiveness and threat in musical passages. *Empirical Musicology Review*, *1*(3), 170-177. https://doi.org/10.18061/1811/24068

Léveillé Gauvin, H. (2018). Drawing listener attention in popular music: Testing five musical features arising from the theory of attention economy. *Musicae Scientiae*, *22*(3), 291-304. https://doi.org/10.1177/1029864917698010

Mohammad, S., & Turney, P. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436-465, 2013. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Morton, E.S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855-869. https://doi.org/10.1086/283219

Morton, E.S. (1994). Sound symbolism and its role in non-human vertebrate communication. In L. Hinton, J. Nichols, & J. Ohala (Eds.), *Sound symbolism* (pp. 348-365). Cambridge, UK: Cambridge University Press.

Paul, B., & Huron, D. (2010). An association between breaking voice and grief-related lyrics in country music. *Empirical Musicology Review*, *5*(2), 27-35. https://doi.org/10.18061/1811/46747

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and Psychometric properties of LIWC2015.* Austin: University of Austin at Texas. Retrieved from https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf?sequence=3

Rigg, M. (1937). Musical expression: an investigation of the theories of Erich Sorantin. *Journal of Experimental Psychology*, *21*(4), 442-455. https://doi.org/10.1037/h0056388

Stouten, J., Gilissen, S., Camps, J., & Tuteleers, C. (2012). Music is what feelings sound like: the role of tonal and atonal music in unethical behavior. *Ethics & Behavior*, *22*(3), 189-195. https://doi.org/10.1080/10508422.2011.644499

Sun, S. H., & Cuthbert, M. S. (2018). Emotion painting: lyric, affect, and musical relationships in a large lead-sheet corpus. *Empirical Musicology Review*, *12*(3-4), 327-348. https://doi.org/10.18061/emr.v12i3-4.5889

Warren, C. S. (2014, October 23). Music is what feelings sound like: music can help us express emotion that is hard to verbalize. *Psychology Today*. Retrieved from https://www.psychologytoday.com/us/blog/naked-truth/201410/music-is-what-feelings-sound