Modeling Musical Complexity: Commentary on Eerola

JOSHUA ALBRECHT
The University of Mary Hardin-Baylor

ABSTRACT: In his paper, “Expectancy violation and information-theoretic models of melodic complexity,” Eerola compares a number of models that correlate musical features of monophonic melodies with participant ratings of perceived melodic complexity. He finds that fairly strong results can be achieved using several different approaches to modeling perceived melodic complexity. The data used in this study are gathered from several previously published studies that use widely different types of melodies, including isochronous folk melodies, isochronous 12-tone rows, and rhythmically complex African folk melodies. This commentary first briefly reviews the article’s method and main findings, then suggests a rethinking of the theoretical framework of the study. Finally, some of the methodological issues of the study are discussed.

Submitted 2016 February 18; accepted 2016 March 5.

KEYWORDS: *complexity, melody, information theory, musical expectancy*

THERE is significant historical overlap between information theoretic approaches to music and theories about the role of musical expectations (Huron, 2006; Meyer, 1956, 1957; Narmour, 1990). As Eerola points out in his article, both information theory and theories of musical expectation can inform Berlyne’s (1971) aesthetic theory, especially in relation to complexity. Given this theoretical overlap, it is not surprising that Eerola uses both information theory and expectancy violation as frameworks for building models of melodic complexity.

The relationship between melodic complexity and expectancy violation is relatively straightforward. Simply stated, the less that a melody progresses according to a listener’s expectations the more complex it is likely to be perceived. Similarly with information theory, the less probable a musical utterance is given some baseline probability – for example, the statistical regularities within a corpus – the higher the information entropy of that utterance and the more complex it is likely to be perceived.

Both of these ideas have been explored in prior work. Eerola and North (2000) tested a model that measured the amount of expectancy violation in melodies using various metrics involving tonal, intervallic, and rhythmic factors. They found that their expectancy-violation model was able to account for 90% of the variance in participants’ ratings of perceived melodic complexity. Pearce and Wiggins (2004) likewise tested a model that measured the information entropy of various monophonic melodies. A type of synthesis between these approaches was offered in Eerola et al. (2006), who identified the eight principles of melodic complexity outlined on pages 4-5 of the reviewed article. In the 2006 study, they found that the eight principles comprising their EV_8 model were significantly correlated with ratings of perceived melodic complexity. Helpfully, the EV_8 model is available through the author’s MIDI toolbox (Eerola & Toiviainen, 2003), a powerful set of analytical tools.

This article fits into the author’s research trajectory as a next step that follows up on much of the important work described in the preceding paragraph. The incredible amount of variance explained by the melodic complexity models in Eerola and North (2000) and Eerola and Toivianinen (2003) is striking, suggesting that the factors involved are likely significant components of perceived melodic complexity. However, in any model-building from existing data, there is always a danger of overfitting. The current article offers an opportunity to test the generalizability of the prior models with new data. Specifically, this article 1) tests the generalizability of the EV_8 model (and some proper subsets of the EV_8) on a number of different datasets compiled from different authors and sources, 2) tests a number of low-level information-theoretic models on the same datasets, and finally 3) offers a critical comparison between the models.

METHODOLOGICAL APPROACH AND MAIN FINDINGS

This article examines and compares a number of models that predict listener evaluations of melodic complexity on several datasets of monophonic melodies. These models use a number of predictors, mostly drawing on structural properties of the melodies that have been previously linked in some way to conceptions of complexity. Eerola distinguishes between these properties, identifying some measures that relate to expectancy violation and some that relate to information theory.

In short, Eerola compares several different models with varying degrees of performance. Model selection, however, is complicated by the fact that model performance is dependent on the dataset. In other words, each model tends to perform well on some datasets that others perform poorly on and vice versa. Another complicating factor is that there is a lot of overlap in what the models measure. For example, the three ‘expectancy-violation’ models share a core of three parameters, with one model adding one extra parameter and the other adding five extra parameters. Additionally, he looks at a model that only measures likelihood of second-order scale-degree transitions, a model that only measures likelihood of second-order interval transitions, and one that combines the two.

Ultimately, Eerola argues for the EV_4 model as the best performer for the current datasets. This model predicts listener evaluations of complexity using the predictors: 1) tonal ambiguity, 2) pitch proximity, 3) entropy of duration distribution, and 4) entropy of pitch-class distribution. This model was able to account for 44.7% of the variance in listener ratings of melodic complexity across all datasets. Although this number is lower than some of the results from Eerola et. al (2006), in which they found that up to 89% of the variance in participant responses could be explained by the EV_{10} model, the numbers are nevertheless quite high for an examination of such different datasets.

What is particularly interesting about Eerola’s study is that it is a meta-analysis of findings from seven different experiments. The experiments span 25 years, several different authors, and represent experiments that were conducted for different reasons, so the variety is particularly rich. Even more noteworthy is the variety of musical materials used, including isochronous 12-tone rows, chromatic modifications of Frere Jacques, and highly syncopated African melodies.

I find this variety of musical materials to be particularly compelling because a common problem arising in complexity studies is the use of stimuli that do not cover a wide enough range of complexity (for a review, see Sluckin, Hargreaves, & Colman, 1982). When this happens, one of two problems might surface. In the first case, participants are likely to compare the melodies with one another, and so participant ratings are likely to be scaled to the dataset they were exposed to. This will magnify differences between melodies that under a more diverse sample would be rated as more similar, resulting in models of complexity that can be overly sensitive to minor differences and that are not generalizable to a different sort of sample.

In the second case, there may not be a wide enough variety between melodies to see the results intended. For example, Berlyne’s (1971) influential aesthetic theory hypothesized that preference for art would follow an inverted-U pattern, in which preference would peak at an optimum level of complexity and trail off as it became too mundane or too complex. One difficulty that experimenters have had in trying to find empirical evidence to support the inverted-U is that a wide range of complexity in stimuli is needed to see this pattern. If the range is too narrow, only the upward or downward slope of the inverted-U might perhaps be seen, and the results might instead suggest a simpler linear relationship between complexity and preference. By opening up his models to such a wide variety of stimuli from different experiments, Eerola is simultaneously minimizing both effects.

The nature of the meta-analysis, however, does not fully avoid the above-mentioned problems. For example, it is important to remember when looking at the data that ratings of complexity are certainly not comparable across datasets. Although Eerola scales each rating between 0 and 1 within each dataset, it must certainly be the case that a rating of 1 for African melodies would mean something entirely different from a rating of 1 for isochronous simple melodies. This effect can be seen in the extreme differences in variance explained between models tested on individual datasets and the combined variance explained of all datasets, which is much lower. Probably the most reliable data would come from re-conducting the study with a new set of participants who hear all the melodies from all the datasets. Of course, while yielding better data, this approach would be costly in both time and money. Given these limitations, it is understandable that Eerola conducts this meta-analysis the way he does. Nevertheless, the fact that he is still able to account for so much overall variance in the participants' ratings is still very impressive.

RETHINKING THE DICHOTOMY

Eerola contrasts models of melodic complexity that are based on "expectancy violations" with those that are "information-theoretic." While this dichotomy carries the primary weight of Eerola's rhetorical strategy in this paper, I believe that it is problematic. Although "the contribution of the individual principles [are] critically evaluated with additional datasets and an attempt [is] made to keep [the two] principles conceptually separate" (p. 6), I'm not sure the distinction has enough traction to support the article's theoretical framework.

Although he frames the models from Eerola et al. (2006) as primarily based on expectancy-violation, there is actually a strong focus on information theory concepts in that article. He acknowledges as much on p. 6 when he says his model perhaps suffers "from overfitting since it was initially developed to predict complexity…" and that some of "the principles contained in this formulation of the [expectancy-violation] model consist of information-theoretic predictors (principles 4-6) that could be argued to be conceptually different from the other principles that are based on averaging indices related to perceptual qualities of notes."

If Eerola's "expectancy violation" models are infused with information theoretic categories, the foundations of information theory itself are also infused with concepts of expectancy violation. One of the primary components of information theory is the concept of information entropy, which measures the uncertainty of a signal (Shannon, 1948). Because uncertainty is directly related to expectation, it is a bit odd that Eerola would conceptually separate the two as being distinct types of models.

Rather, the language of uncertainty, which is focused on the signal, and the language of expectation, which is focused on the perceiver, refer to much the same thing from different perspectives. Eerola captures this problem succinctly when he says, "[t]he difference between the types of model is theoretically notable but in practical terms subtle, since the rule-based principles are probably heuristics that capture statistical regularities in the music, and vice versa" (p. 16).

I believe that instead what is happening is that the two different frameworks in this paper are focused on different *types* of expectations. The two different types of expectations are similar to what Huron (2006) calls "dynamic" and "schematic" expectations or what Narmour (1990) calls "intraopus" and "extraopus" expectations. Intraopus or dynamic expectations are derived as a melody (in this case) is listened to. This can be modeled by examining what the melody *actually does* and using that to establish a set of expectations. Taking the author's category 7 as an example, a greater amount of rhythmic variation in a given melody would mean that any given rhythmic value would be hard to predict, increasing information entropy and perceived complexity. It is this dynamic level of complexity that I believe the author is focusing on with his EV models.

Conversely, extraopus or schematic expectations are derived from knowledge of a given repertoire. For example, knowing that in a given repertoire leading-tones usually move to either scale degrees 1, 2, or 5 would build a set of expectations for when the leading-tone is heard. By violating the statistical properties of the style that the melody is derived from, the melody is heard as more complex or could be conceived of as having a higher amount of information entropy. I believe that this is the primary element being modeled in what the author calls "information-theoretic" models.

Of course, both types of expectations are at play in any encounter with a specific melody. Taking the above example, if a melody deceptively resolves all of its leading-tones, it is reasonable to suppose that a listener enculturated in Western classical music may start to expect this over time. Although there may still be some measure of surprise as a result of the schematic expectations of the style, one may expect that the work would become gradually more predictable over time and complexity ratings would gradually decrease.

In other words, both kinds of models in this article could properly be said to model information entropy or musical expectations, depending on the frame of reference, but the first kind of model focuses on dynamic expectations derived from a stand-alone melody whereas the second kind of model focuses on schematic expectations derived from a particular instance of a melody from a broader style (in this case, the Essen folksong collection).

As a thought experiment, it is interesting to wonder what would have happened in the seven experiments if the instructions were altered to read "how much does this violate your musical expectations?" or "how much information do you think is being communicated by this?" instead of "how complex is this?". My guess is that while there may be some differences in the ratings, which themselves would be interesting to examine, the excerpts would likely be rank-ordered in very similar ways. If this were the case, it would be consistent with the notion that perceptions of 'degree of expectedness' and 'amount of information communicated' account for roughly the same perceptual categories as each other and the more intuitive perception of 'melodic complexity.'

SOME METHODOLOGICAL CONSIDERATIONS

The Dangers of Overfitting

As mentioned earlier, any time regression models are built to account for variance within a dataset with a set of predictors, there is always a danger of overfitting to that dataset. In this article, Eerola uses some statistical methods to mitigate this problem. Specifically, by using a number of k-fold cross validations (p. 9) it is likely that he was able to model the real variables that contributed to variance in participant ratings within his datasets.

Nevertheless, finding the variables that explain perceived complexity within these datasets is not the same as finding the variables that explain perceived complexity within melodies in general. The melodies tested were primarily folk melodies, and many were isochronous, with some 12-tone rows, new age melodies, and African syncopated melodies included. The form of the models or their ability to predict future ratings of perceived complexity might be quite different if instead Eerola had used classical themes, opera arias, Gregorian chants, or Tin Pan Alley songs.

Also, because the datasets are sampled from different 'populations' of melodies and are not the same size, the resulting models will tend to amplify differences in complexity in the larger datasets. For example, the folk melodies from the Essen collection (D6) make up a full 25% of the total 205 melodies. The 12-tone rows (D3) account for less than 10% of the melodies and are of such a different type that they are likely accounted as outliers in the models. Tables 6 and 7 reveal that none of the models can account for any more than 28% of the variance in that dataset, and that the median can only account for 11% of the variance.

It is worth recalling here that this article likely represents efforts to generalize the models from Eerola et al (2006) to different datasets. By looking at new datasets and reducing the models to fewer variables, the paper helps to determine which elements of the models were overfit to the prior datasets. Still, even the models chosen for the final analysis in Table 7 were picked from a large number of trial models based on their fit with the data (especially the 'information-theoretic' models on pp. 8-9). As a result, overfitting is still a strong possibility; the difference is that any overfitting is happening with a larger dataset.

Within- vs. Between-Groups

In this study, Eerola is trying to predict ratings of perceived melodic complexity, but all of these ratings were collected within each of the individual studies. Although the scores have been normalized within groups to scores ranging from 0-1, they are not really comparable between datasets, as mentioned above. Additionally, there is likely an uneven distribution of complexity in the datasets. It is likely, for example, that there is greater variance of melodic complexity in the African folk melodies than in the isochronous 12-tone rows. As a result, the ratings are not absolute, but are in relation to the other melodies within each dataset. Of course, this problem is completely unavoidable without re-running the experiment with new participants and simultaneously testing all the datasets.

Nevertheless, Eerola builds models with combinations of all of the ratings. Essentially, this creates an extra source of error. For example, the least complex melody from the D7 dataset is likely still more rhythmically complex than all other melodies, but it will have a low rating because it is less complex than the rest of that dataset. As a result, the model will underestimate the complexity of most of these least complex D7 melodies. However, it will overestimate the complexity of the most complex melodies, because D7 only represents 21% of the dataset. The problem is visualized in Figure 2 in the article.

The fact that Eerola is able to so accurately predict ratings of complexity despite this extra source of error is indeed evidence that these models "already have demonstrated robustness" (p. 3). However, although he reports adjusted R^2 values for the expectancy violation models, he only reports correlations of the information-theoretic models within groups. Because the model is built from all datasets as a whole, it would be appropriate to compare the within-group correlations with the between-group correlations for all of the models. Non-adjusted R^2 values are provided in Table 1.

As can be seen in Table 1, the best performing model remains EV_4, which is still able to account for a fairly impressive 46% of the variance in participant ratings of complexity across all datasets. While understandably losing some predictive power due to using one fewer predictor, EV_3 can still explain a substantial 42% of the variance across all datasets. However, all of the models perform less well across all of the datasets. This is not surprising, given the extra error involved in the normalized ratings. Notably, the biggest declines in explanatory power are the two models that rely on extremely low-level second-order pitch and interval information, PC_3 and IV_3. These models only account for 9% and 15% of the variance across all the datasets.

It is likely that these numbers are artificially deflated because of the extra error involved in collapsing several different studies with different instructions and different datasets into the same models. However, the lower numbers do raise questions about the generalizability of the models to other melodies. Future research testing these models on perceptual ratings of complexity on a new sample of melodies would help to untangle some of these questions. If the models below are tested without any new selection or modification, it would also shed light on how much overfitting there is in the models.

Table 1. A comparison of the within-group median correlation and variance explained (provided in Table 7 of the article) with the between-group correlations and variance explained across all datasets. As expected, between-group performance is not as strong.

	Within-group correlation median	*Within-group% variance explained (R^2) median*	*Between-group correlation*	*Between-group% variance explained (R^2)*
EV_8	.46	21%	.36	13%
EV_4	.74	55%	.68	46%
EV_3	.68	46%	.64	42%
PC_3	.72	52%	.30	9%
IV_3	.65	42%	.39	15%
IT_2	.69	48%	.58	34%

Key Estimation

The PC models (and by extension IT_2) that Eerola uses rely on scale-degree transitions, and so a reliable key is needed for each melody to determine which scale degrees the pitch class transitions represent. To calculate the scale-degree transitions in his melodies, Eerola uses the Krumhansl-Schmuckler key-finding algorithm (Krumhansl, 1990). Because these models rely on schematic or extra-opus information, a large reference corpus is needed from which to derive statistical distributions of scale-degree transitions. Eerola uses the Essen folk music corpus for this purpose (Schaffrath, 1995).

One benefit of the Essen database is that each melody is already assigned a key by an analyst. Although it is clear that Eerola uses the Krumhansl-Schmuckler key-finding algorithm to classify the scale-degree transitions in his model melodies, it is not clear whether he uses the same method to determine the scale-degree transitions for the entire Essen database or whether he uses the assigned keys in the database.

What is clear is that the Krumhansl-Schmuckler algorithm suffers from some systematic inaccuracies. In a study of accuracy rates for various key-finding algorithms, Albrecht and Shanahan (2013) found that the Krumhansl-Schmuckler algorithm was only 74.2% accurate in assigning keys to a large collection of 957 common-practice era classical music. The algorithm actually fares worse on folk melodies, with only 68.7% accuracy (seen in Table 2). Even in Eerola's article, his *Melody B* (Figure 1) is incorrectly assigned the key of E minor. Such low key estimation accuracy rates raise questions about the usefulness of the PC_3 model, as Eerola acknowledges (p. 7, p. 13).

The Albrecht and Shanahan (2013) key-finding algorithm performed significantly better on the dataset of 957 classical works, with an accuracy of 93.1% ($p < .0001$). Using this algorithm, accuracy rates on the Essen database were improved to 87.7% (Table 2). Another algorithm worth considering for this dataset is the Aarden-Essen algorithm (Aarden 2003). This algorithm benefits from having been trained on the Essen database itself. However, accuracy rates are only a marginally better 87.9% on this database (p = .81). Albrecht and Shanahan (2013) proposed a meta-algorithm that combines the strengths of their proposed algorithm and the Aarden-Essen algorithm. That meta-algorithm performs significantly better than any of the others (p < .0001) at 92.0% accuracy.

The low amount of between-group variance accounted for by PC_3 and IT_2 (9% and 34%, respectively) might be in part because of incorrect key estimation. If only 69% of melodies are correctly assigned keys, the algorithm may assign less common second-order scale-degree transitions to the melody than the listeners, who presumably hear them in their correct key-context. By using a more accurate key-finding method, or having the keys provided by a skilled analyst, these models may produce closer fits to the perceptual data.

What is perhaps most telling is that the information-theoretic models, which were trained on the statistical regularities within the Essen dataset, performed worse on the Essen dataset (D6) than on most of the others. The exception, however, is the syncopated African folk melodies and the 12-tone rows; in these cases, most of the variance in perceived melodic complexity is likely due to the syncopations and rhythms and in the dissonances of the leaps, respectively. As Table 6 reveals, the PC models performed notably worse than the IV models for D6. It is possible this loss of performance further indicates the problem with key-finding.

Table 2. Accuracy rates of four different key-finding algorithms on the database of 6,210 folk melodies in the Essen database. The best performing algorithm was the Albrecht and Shanahan (2013) meta-algorithm. All differences were significant (p < .0001), except for the difference between the Albrecht and Shanahan (2013) and Aarden-Essen (2003) algorithms (p = .81).

Algorithm	*Accuracy rates on 6,210 Essen folk melodies*
Krumhansl-Schmuckler (1990)	68.7%
Aarden-Essen (2003)	87.9%
Albrecht and Shanahan (2013)	87.7%
Albrecht and Shanahan (2013) meta-algorithm	92.0%

CONCLUSION

Even with the above methodological caveats, the Eerola article is important for extending work on previous models of melodic complexity. His results show that a relatively large portion of the variance in listener ratings of melodic complexity can be modeled with just a few low-level parameters measuring expectancy violation. Most significantly, this is true even using different datasets of melodies, in experiments conducted by different researchers, with different types of participants, and with different instructions.

However, there are still several unanswered questions. The success of the models Eerola tested may in part be due to overfitting. To what extent these models are overfit to the data (even with the large number of different datasets) is unclear. Newly gathered data on differently designed studies with a large range of melodies can help to shed light on this question. Opening the range of participants to more cultures, broader age ranges, more diverse levels of musical background, and broader socioeconomic backgrounds can also bolster the claims of generalizability for these models.

Additionally, monophonic melodies, though a central element of many folk music traditions, are relatively simple compared to homophonic or complex polyphonic music. Even if the models presented in this article are close to truly measuring melodic complexity in monophonic melodies there is still a lot of work to do to understand how these parameters fit into more complex musical textures. What happens when a melody fits loosely with or even conflicts with a harmony? What is the interaction between perceived complexity and the relationship between text and music?

Though there is much yet to be done surrounding the question of musical complexity, studies like the current one give researchers much to be optimistic about. When the same melodic elements are reliably correlated with perceived complexity across varied studies, there is what seems to be converging evidence on new theories of melodic complexity. Nevertheless, there are many further avenues to investigate, and it seems likely that pursuing these research lines would undoubtedly reveal new and interesting complexities to this story.

REFERENCES

Aarden, B. (2003). *Dynamic melodic expectancy.* PhD Dissertation (Columbus: Ohio State University).

Berlyne, D. E. (1971). *Aesthetics and psychobiology.* New York, NY: Appleton Century Crofts.

Eerola, T., & North, A. C. (2000). Expectancy-based model of melodic complexity. In C. Woods, G. B. Luck, R. Brochard, S. A. O'Neill, & J. A. Sloboda (Eds.), *Proceedings of the sixth international conference on music perception and cognition* (pp. 1177–1183). Keele, Staffordshire, UK: Department of Psychology, Keele University.

Eerola, T., & Toiviainen, P. (2003). *MIDI toolbox: MATLAB tools for music research*. Jyväskylä, Finland: University of Jyväskylä.

Eerola, T., Himberg, T., Toiviainen, P., & Louhivuori, J. (2006). Perceived complexity of Western and African folk melodies by Western and African listeners. *Psychology of Music*, *34*(3), 341–375. http://dx.doi.org/10.1177/0305735606064842

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception, 19*(1), 1–64. http://dx.doi.org/10.1525/mp.2001.19.1.1

Krumhansl, C. L. (1990). *The cognitive foundations of musical pitch.* New York: Oxford University Press.

Meyer, L. B. (1956). *Emotion and meaning in music.* Chicago, IL: University of Chicago Press.

Meyer, L. B. (1967). *Music, the arts, and ideas.* Chicago, IL: University of Chicago Press.

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model.* Chicago, IL: University of Chicago Press.

Pearce, M. T., & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*(4), 367–385. http://dx.doi.org/10.1080/0929821052000343840

Schaffrath, H. (1995). The Essen folksong collection. In D. Huron (Ed.), *Database containing 6,255 folksong transcriptions in the Kern format and a 34-page research guide [computer database]*. Menlo Park, CA: CCARH.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, Illinois: University of Illinois Press.

Sluckin, W., Hargreaves, D.J., & Colman, A.M. (1983). Novelty and human aesthetic preferences. In J. Archer& L. Birke (Eds.), *Exploration in animals and humans* (pp. 245-269). Workingham: Van Nostrand Reinhold.