# Computational Models of Temporal Expectations

Lauren K. Fink[1,2]
[1] Department of Music, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany
[2] Max Planck – NYU Center for Language, Music, and Emotion (CLaME), New York, USA
lauren.fink@ae.mpg.de

## Abstract

With Western, tonal music, the expectedness of any given note or chord can be estimated using various methodologies, from perceptual distance to information content. However, in the realm of rhythm and meter, the same sort of predictive capability is lacking. To date, most computational models have focused on predicting meter (a global cognitive framework for listening), rather than fluctuations in metric attention or expectations at each moment in time. This theoretical contribution reviews existing models, noting current capabilities and outlining necessities for future work.

KEYWORDS: *rhythm, meter, percussion, salience, prediction error, novelty, surprise*

## Introduction

One major goal of music cognition research is to understand how certain properties of an acoustic signal give rise to specific physiological, psychological, and behavioral responses. Such an endeavor is not straightforward, as the body and nervous system filter the incoming acoustic signal, suppressing certain aspects, amplifying others, and creating emergent percepts dependent on different temporal windows of integration. While many mysteries remain regarding the transformation from acoustic signal to psychological percept or subjective experience, numerous strides have been made, particularly in the realm of predicting listeners' tonal or harmonic expectations (e.g., Janata et al., 2002; Pearce & Wiggins, 2012; Cancino-Chacón, Grachten, & Agres, 2017).

While most Western music does organize itself around a tonal center – allowing for specific harmony-based expectations – plenty of music does not. Because one ultimate goal of the field should be a model of expectancy that can operate on never-before-heard music, especially from styles outside of Western, dominant culture (Jacoby et al., 2020; Baker et al., 2020), I focus here on the case of percussive music. Independent of musical culture or style, percussive music may or may not be pitch-based. Hence, I ask which computational models can provide continuous, meaningful predictions of listeners' physiological responses or psychological expectations while listening to a drum set solo, a tabla performance, a dùndún ensemble, a marching drumline, or pieces like John Cage's *Credo in US*?

At present, there exist a variety of impressive models to extract the beat or meter from a given piece (e.g., Klapuri, Eronen, & Astola, 2006; Volk, 2008; Smith & Honig, 2008; Tomic & Janata, 2008; Temperley, 2010; Grosche & Muller, 2010; Large, Herrero, & Valesco, 2015; van der Weij, Pearce, & Honing, 2017; Boenn, 2018; Lartillot & Grandjean, 2019), with annual competitions held (e.g., MIREX Audio Beat Tracking Competition; Holzapfel et al., 2012), and algorithms constantly improving. While beat and meter provide an important foundation for when listeners might perceive certain moments to be more salient than others, they are not necessarily clearly present in every style of music. Also, beat and meter can fluctuate over time, such that one may need to conceive of salience within different metric frameworks of listening at different points in time (an approach exemplified via theoretical analysis in London's, 2012, *Hearing in* Time, Ch. 7).

Can shifting metric frameworks of listening be modelled online in a manner that allows for updating predictions, especially as a function of repeatedly violated expectations? How might we account for the likelihood of certain perceived meters over others, given the number of present pulse layers (see Gotham, 2015, for discussion)? Can our models incorporate asymmetries in perception with regard to increases vs. decreases in certain features (e.g., regularity, loudness, etc.)? What about the role of listener familiarity and enculturation in predisposing a listener to certain tempo-metrical types (London, 2012; London, Polak, Jacoby, 2016; Holzapfel, 2015)? Or the role of repeated exposures in shaping the temporal context (e.g., local vs. global) over which listeners are attending (Margulis, 2014)? Finally, how might we model temporal expectations when no clear beat, meter, or typical tonal relationships exist?

This short article is not meant to be an exhaustive review of state-of-the-art models, but rather an entry

point for interested researchers and students. Below, I provide an overview of relevant criteria when considering models of musical expectation more broadly. I then briefly outline a few models possibly able to meet the previously defined task of generating meaningful predictions in the case of percussive music. In so doing, I simultaneously highlight strengths, limitations, and directions for future modeling efforts.

## Model Considerations

*Biological Plausibility*

One important note at the outset is that not all models of musical expectations aim to be biologically plausible. While biologically plausible models of auditory processing, from the cochlea to the cortex, do underpin some existing models of various aspects of music cognition (e.g., Collins et al., 2014; Large, Herrera, Velasco, 2015), biological plausibility is not necessarily a requirement (i.e., one could build a model that accurately predicts human perception or behavior by using signal transformations very different from those thought to be employed by the nervous system). In general, the validation of computations on a neural level is likely relegated to, or informed by, research in animal models (e.g., Elie & Theunissen, 2019) and/or computational neuroscience (e.g., Tomková, 2015; Masquelier, 2018). Whether biological plausibility matters depends on one's specific research question and context.

*Mathematical Framework*

Related to, but not synonymous with the notion of biological plausibility, is the mathematical framework in which a model of musical expectations operates. The current leading models come from a variety of mathematical stances, from probabilistic frameworks to non-linear oscillatory ones. These approaches reflect more than just a mathematical means to an end; they often signify the fundamental concept and perhaps even level of explanation at play in the model.

Oscillatory models take a bottom-up approach, closely aligned with Dynamic Attending Theory (Large & Kolen, 1994; Large & Jones, 1999). Specifically, the idea is that periodicities (oscillations) in music can entrain oscillations of the nervous system (bottom-up). In some models, the predicted perceived meter (as determined by the most active oscillators) can then direct attention to the incoming stimulus (see e.g., Hurley, Fink, Janata, 2018). Such models have a sort of short-term memory built in, in the form of the decay time of the oscillators, but do not (at present) integrate over longer time scales in the way that Bayesian or Markov models can. Nonetheless, enculturation can be built into oscillatory models in the form of pre-set couplings between oscillators of different frequencies, which reflect the metric norms of the listeners' culture, as in the model of Large, Herrera, and Valesco (2015).

Alternatively, prior experience can be modelled in the form of a Bayesian prior. As such, Bayesian models can easily account for accumulated sensory evidence over time (at short and long time scales; see, e.g., Skerritt-Davis & Elhilali, 2018), as well as listener enculturation (see e.g., van der Weij, Pearce, & Honing, 2017). Bayesian approaches fit well within predictive coding theory (Friston, 2002; Vuust & Witek, 2014), which is compatible with Dynamic Attending Theory, and reflect the idea that past experience is used to generate predictions about incoming stimuli. The mismatch between what was expected and what occurs is used to update a probability distribution and thereby modify future predictions. Some Bayesian models also include parameters to reflect human memory, as well as observation noise (e.g., Skerritt-Davis & Elhilali, 2018).

*Input Signal*

Many leading models of melodic (e.g., Pearce, 2005), harmonic (e.g., Harrison, Bianco, Chait, Pearce, 2020), and temporal expectations (e.g., Forth et al., 2016) or meter (e.g., Large, Herrera, & Velasco, 2015) operate on symbolic (e.g., **kern, MIDI, etc.) data. Databases of music meticulously converted into such formats exist (e.g., http://kernscores.stanford.edu/), which allow users of models requiring them to test their predictions on these specific corpora. However, these symbolic representations require transcription or automated conversation, which costs additional time and is subject to inaccuracies. Further, by reducing music to constituent features, some critical information that drives human perception may be lost (i.e., features are treated discretely, rather than continuously; aspects of timbre and expressivity cannot be represented). Most important for the present account, unless one is studying tonal music played on mallet instruments, percussive music is not so simply translated into discrete alphabets, as it might involve different ways of striking or preparing the same instrument to generate subtly different timbres, the incorporation of human voice, pitch bends, water gongs, and so on. Thus, a model that can operate on raw audio files, or continuously extracted acoustic features, would be ideal.

One additional issue with regard to input signal is not the data file format, but rather the amount of training data required to produce an estimate of expectation over time. Some models require training on a corpus of music to statistically learn the syntax of particular musical styles, while others can operate only a single piece in a purely bottom-up manner. Still others can operate on single pieces (taking into account only what has been heard of the piece so far) or use a previously trained model for defining initial probabilities. Again, whether these aspects are important, depends on one's question (e.g., modeling a naïve vs. well-experienced listener), and music of interest (e.g., monophonic MIDI representations from a style with copious examples vs. percussive music from a style which may not be represented by available corpora).

*Salience, Information Content, Surprisal, Novelty*

While each of the words in the above heading have slightly different meanings, especially when considered with respect to the mathematical equations they represent, their underlying basic concept is the same: a continuous (symbol-wise, or sample-wise) prediction of the allocation of a listeners' attention or expectations.

In some cases, however, such words may index slightly different constructs. For instance, in the model utilized in Hurley et al. (2018), high values of *salience* indicate a prediction of increased attention to that moment in time. On the other hand, in the model of Skerritt-Davis & Elhilali (2018) high values of *surprisal* indicate moments when expectations have been violated. *Information content* (e.g., as implemented in van der Weij, 2017) indexes a construct synonymous to surprisal (i.e., an unexpected event likely to elicit a prediction error). All of these measures, in essence, should be correlated, in that moments of expectation violation re-orient attention. However, moments of high salience could be expected (as a culmination of the preceding context) and do not necessarily evoke an event-related potential in physiological data, as would be expected by moments of high surprisal.

Taking a slightly different, but related, semantic stance, the *novelty* metric (Lartillot, Cereghetti, Eliard & Grandjean, 2013) from the music information retrieval (MIR) toolbox (Lartillot & Toiviainen, 2007), is not so much grounded in ideas about listeners' attention but rather in notions of segmentation (i.e., how listeners might parse an acoustic signal into different sections). *Novelty* thus represents how likely one is to perceive a transition at any particular moment. This construct is related to attention / expectation, but will likely show a somewhat different relationship to physiological and behavioral activity than the previously defined measures.

## Models of Temporal Expectations

I now turn to the task set forward at the beginning of this article: a model that can provide a continuous prediction of listener expectation over time for a recording of percussive music. Without finding some way to transcribe such a recording into a symbolic format, we are left with few options. I will briefly summarize three relevant models that can, in theory, accomplish the task, but, in practice, may require additional validation.

From a bottom-up, oscillatory framework, the extension of the Beyond the Beat model (Tomic & Janata, 2008) validated in Hurley et al. (2018) and Fink et al. (2018) functions in the current context. Raw audio can be input to generate a prediction of salience over time. The prediction is calculated through a series of steps involving a cochlear model, amplitude envelope extraction and onset detection in multiple frequency bands, feeding onsets through a bank of oscillators (tuned to frequencies up to 10 Hz and exhibiting decay), and averaging the activity of only the most active oscillators at each point in time. This model has been shown to predict fluctuations in listeners' attention, measured via perceptual thresholds (Hurley et al., 2018) and pupil size (Fink et al., 2018), but still requires further validation against a broader range of stimuli and against other models (discussed below).

From the realm of music information retrieval, *mirnovelty* (MIR toolbox) may be useful in the current context, as it reflects changes in the high-level structure of music, by taking into account contrasts in temporal scale on an instant-by-instant basis. It can be calculated in a kernel-based or multi-granular (Lartillot et al., 2013) way, from the similarity matrix of the acoustic signal or particular acoustic features (i.e., by taking into account one, or many different, temporal windows of integration). Given that this measure indexes larger-scale moments of structural change (i.e., the probability and importance of transitions between sections), one might expect behavioral or physiological measures related to "chunking" to most closely follow this prediction, though extensive testing against human behavior (Hartmann, Lartillot, & Toiviainen, 2017) and/or physiology has yet to be conducted.

From a Bayesian probabilistic framework, the Dynamic Regularity Extraction (D-REX) model of Skerritt-Davis & Elhilali (2018) predicts changes in the underlying statistical structure of a stimulus. Its

predictions can be based only on what has been processed for the current stimulus or informed by a previously trained model. D-REX builds probabilities about current events based on 1) the current context and 2) the probability that the context (underlying statistics) has changed. The model incorporates both a memory and noise parameter to more accurately simulate human perception. Input to the model is any continuous acoustic feature of choice; output is prediction error, or *surprisal*, over time. D-REX predictions have been validated against electroencephalographic data using stochastic, random, and regular monophonic stimuli. To date, only predictions based on pitch have been employed (see Skerritt-Davis & Elhilali, 2018; 2019), though predictions based on other acoustic features, or even combined over multiple features is possible.

## Discussion

At present, few models exist that can make continuous predictions of listener expectations from continuous audio stimuli. A systematic comparison of all relevant models, using the same stimuli, participants, and physiological experiments is required. It would also be ideal to also include symbolic models (e.g. Harrison et al., 2020) in such a comparison. However, this task is not simple, as equivalence needs to be achieved regarding discretized symbols vs. continuous acoustic features and model outputs. It may be the case that such comparisons are not so meaningful and that a more fruitful approach involves characterizing the particular research questions and musical styles for which each model is most well-suited. Future work along such lines will move the field forward.

Additionally, future models might incorporate deep learning approaches. At present, such models are used to generate novel music from training on tokens (e.g., MuseNet; Payne, 2019) or, impressively, on raw audio (e.g., WaveNet, Oord et al., 2016; Jukebox, Dhariwal et al., 2020). These models could be adapted to generate predictions about what will happen next in a given piece of music, rather than to generate novel music.

## Conclusion

Inspiring computational advances have been made in the past decades to predict listeners' expectations while engaging with music. Future advances may involve:

- Generating predictions across multiple time scales, acoustic features, and perceptual classes (e.g., streams, textures).
- Pooling temporal expectations across multiple modalities. Musical engagement often involves a strong visual component, which can also shape expectations. Perhaps models of acoustic and visual saliency could be combined to study expectations in audiovisual contexts.
- Filtering predictions based on physiological limitations and processing asymmetries.
- Tuning predictions based on listener familiarity (both in terms of exposure to the particular stimulus in question, as well as the cultural context of the stimulus and listener).

## Acknowledgements

I thank Lea Fink, Angela Nazarian, and Lindsay Warrenburg for helpful comments on an earlier version of this manuscript.

## References

Baker, D. J., Belfi, A., Creel, S., Grahn, J., Hannon, E., Loui, P., ... & Vuvan, D. T. (2020). Embracing Anti-Racist Practices in the Music Perception and Cognition Community.

Boenn, G. (2018). *Computational Models of Rhythm and Meter* (pp. 1-187). Springer.

Cancino-Chacón, C., Grachten, M., & Agres, K. (2017). From Bach to the Beatles: The simulation of human tonal expectation using ecologically-trained predictive models. *arXiv preprint arXiv*:1707.06231.

Collins, T., Tillmann, B., Barrett, F. S., Delbé, C., & Janata, P. (2014). A combined model of sensory and cognitive representations underlying tonal expectations in music: From audio signals to behavior. *Psychological review, 121*(1), 33.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Elie, J. E., & Theunissen, F. E. (2019). Invariant neural responses for sensory categories revealed by the time-varying information for communication calls. *PLoS computational biology, 15*(9), e1006698.

Forth, J., Agres, K., Purver, M., & Wiggins, G. A. (2016). Entraining IDyOT: Timing in the Information Dynamics of Thinking. *Front Psychol, 7*, 1575.

Grosche, P., & Muller, M. (2010). Extracting predominant local pulse information from music recordings. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(6), 1688-1701.

Holzapfel, A. (2015). Relation between surface rhythm and rhythmic modes in Turkish Makam music. *J. New Music Res. 44*, 25–38.

Holzapfel, A., Davies, M. E., Zapata, J. R., Oliveira, J. L., & Gouyon, F. (2012). Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(9), 2539-2548.

Harrison, P. M., Bianco, R., Chait, M., & Pearce, M. T. (2020). PPM-Decay: A Computational Model of Auditory Prediction with Memory Decay. *PLoS computational biology*.

Hartmann, M., Lartillot, O., & Toiviainen, P. (2017). Musical feature and novelty curve characterizations as predictors of segmentation accuracy. In *Proceedings of the Sound and Music Computing Conferences,* Aalto-yliopisto.

Hurley, B., Fink, L., & Janata, P. (2018). Mapping the dynamic allocation of attention in musical patterns. J*ournal of Experimental Psychology: Human Perception & Performance, 44(11), 1694-1711*.

Jacoby, N., Margulis, E. H., Clayton, M., Hannon, E., Honing, H., Iversen, J., ... & Wald-Fuhrmann, M. (2020). Cross-cultural work in music cognition: Challenges, insights, and recommendations. *Music Perception*, *37*(3), 185-195.

Janata, P., Birk, J. L., Van Horn, J. D., Leman, M., Tillmann, B., & Bharucha, J. J. (2002). The cortical topography of tonal structures underlying Western music. *Science, 298*(5601), 2167-2170.

Klapuri, A., Eronen, A., & Astola, J. T. (2006). Analysis of the meter of acoustic musical signals. *IEEE Trans. Audio, Speech, Lang. Process. 14*, 342–355.

Large, E. W., Herrera, J. A., & Velasco, M. J. (2015). Neural networks for beat perception in musical rhythm. *Frontiers in systems neuroscience, 9*, 159.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: how people track time-varying events. *Psychol. Rev. 106*, 119–159.

Large, E. W., and Kolen, J. F. (1994). Resonance and the perception of musical meter. *Conn. Sci. 6*, 177–208.

Lartillot, O., & Grandjean, D. (2019). Tempo and metrical analysis by tracking multiple metrical levels using autocorrelation. *Applied Sciences, 9*(23), 5121.

Lartillot, O., & Toiviainen, P. (2007). "A Matlab Toolbox for musical feature extraction from audio." *International Conference on Digital Audio Effects, Bordeaux*.

London, J., Polak, R., and Jacoby, N. (2016). Rhythm histograms and musical meter: a corpus study of Malian percussion music. *Psychon. Bull. Rev.* 24, 474–480. doi: 10.3758/s13423-016-1093-7

London, J. (2012). *Hearing in time: Psychological aspects of musical meter*. Oxford University Press.

Margulis, E. H. (2014). *On repeat: How music plays the mind*. Oxford University Press.

Masquelier T. (2018). STDP allows close-to-optimal spatiotemporal spike pattern detection by single coincidence detector neurons. *Neuroscience, 389*, 133–140.

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv*: 1609.03499.

Payne, Christine. "MuseNet." OpenAI, 25 Apr. 2019, openai.com/blog/musenet

Pearce, M. T. (2005). The construction and evaluation of statistical models of melodic structure in music perception and composition (Doctoral dissertation, City University London).

Pearce, M. T., & Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Topics in cognitive science, 4*(4), 625-652.

Skerritt-Davis, B., & Elhilali, M. (2018). Detecting change in stochastic sound sequences. *PLoS computational biology, 14*(5), e1006162.

Skerritt-Davis, B., & Elhilali, M. (2019). A model for statistical regularity extraction from dynamic sounds. *Acta Acustica united with Acustica, 105*(1), 1-4.

Smith, L., & Honig, H. (2008). Time–frequency representation of musical rhythm by continuous wavelets. *Journal of Mathematics and Music, 2*(2), 81–97.

Temperley, D. (2010). Modeling common-practice rhythm. *Music Perception, 27*(5), 355-376.

Tomic, S. T., & Janata, P. (2008). Beyond the beat: Modeling metric structure in music and performance. *The Journal of the Acoustical Society of America, 124*(6), 4024-4041.

Tomková, M., Tomek, J., Novák, O., Zelenka, O., Syka, J., & Brom, C. (2015). Formation and disruption of tonotopy in a large-scale model of the auditory cortex. *Journal of computational neuroscience, 39*(2), 131–153.

van der Weij, B., Pearce, M. T., & Honing, H. (2017). A probabilistic model of meter perception: Simulating enculturation. *Frontiers in psychology, 8*, 824.

Volk, A. (2008). The study of syncopation using inner metric analysis: Linking theoretical and experimental analysis of metre in music. *Journal of New Music Research, 37*(4), 259-273.

Vuust, P., & Witek, M. A. (2014). Rhythmic complexity and predictive coding: a novel approach to modeling rhythm and meter perception in music. *Frontiers in psychology, 5*, 1111.