Modelling Perception of Structure and Affect in Music: Spectral Centroid and Wishart's *Red Bird*

Roger T. Dean
MARCS Auditory Laboratories, University of Western Sydney, Australia

Freya Bailes
MARCS Auditory Laboratories, University of Western Sydney, Australia

ABSTRACT: Pearce (2011) provides a positive and interesting response to our article on time series analysis of the influences of acoustic properties on real-time perception of structure and affect in a section of Trevor Wishart's *Red Bird* (Dean & Bailes, 2010). We address the following topics raised in the response and our paper. First, we analyse in depth the possible influence of spectral centroid, a timbral feature of the acoustic stream distinct from the high level general parameter we used initially, spectral flatness. We find that spectral centroid, like spectral flatness, is not a powerful predictor of real-time responses, though it does show some features that encourage its continued consideration. Second, we discuss further the issue of studying both individual responses, and as in our paper, group averaged responses. We show that a multivariate Vector Autoregression model handles the grand average series quite similarly to those of individual members of our participant groups, and we analyse this in greater detail with a wide range of approaches in work which is in press and continuing. Lastly, we discuss the nature and intent of computational modelling of cognition using acoustic and music- or information theoretic data streams as predictors, and how the music- or information theoretic approaches may be applied to electroacoustic music, which is 'sound-based' rather than note-centred like Western classical music.

Submitted 2011 September 30; accepted 2011 October 10.

KEYWORDS: time series analysis; musical structure; musical affect; information theoretic analysis; computational modelling of cognition; electroacoustic music.

PEARCE (2011) supports our focus on studying real-time responses to music, and appreciates our introducing the methods of time series analysis (which have been rarely used in music studies) and using computer-mediated electroacoustic music as part of this analysis. Indeed, in our ongoing work, we make a point of contrasting and comparing responses to the 'sound-centered' electroacoustic musics, as for example Landy (2009) characterizes them, with those to 'note-centered' music, such as piano music from the Western classical music tradition. Our work allows demonstration not only of the predictive capacity of acoustic information streams for perception of musical structure and affect, but also of interactions between individual perceptual/cognitive responses, and their autoregressive properties.

In this article, we respond to three main topics raised by Pearce's comments and by our work to date: the possible influence of the acoustic parameter spectral centroid on perceptions of musical structure and affect; the comparison between group average perceptual responses and individual responses; and finally the nature of computational cognitive modelling and possibilities for developing information theoretic aspects of it in relation to electroacoustic music.

THE POSSIBLE ROLES OF SPECTRAL CENTROID AS A PREDICTOR OF PERCEPTUAL RESPONSES TO *RED BIRD*

In Dean and Bailes (2010), we chose to use spectral flatness as our physical measure of timbre, and found that it was not successfully able to predict listener perceptions of the music. In answer to Pearce's comments (Pearce 2011), we now examine the more commonplace alternative measure of timbre, namely spectral centroid, as a possible predictor of listener perceptions. We measured the spectral centroid of

Wishart's *Red Bird* extract using the MAX/MSP object written by Ted Apel, John Puterbaugh, and David Zicarelli. The fast Fourier transform was computed over 4096 samples (sampling rate of the audio file 44.1K), with a hop size of 4096. The measured spectral centroid value in Hz was then averaged over successive 0.5 s windows. As previously, time series models were developed on the basis of parsimony though here using the more stringent Bayesian Information Criterion (BIC; which penalises more for parameters). All models mentioned below which give what we refer to as significant interpretations, also gave rise to white noise residuals in the modelled parameter(s) under discussion.

We first assessed Granger Causality in bivariate analyses, where one perceptual stream was modelled on the basis of vector autoregression and one acoustic variable, spectral centroid. Both variables were treated as endogenous. For perception of change or dchange (as in Dean & Bailes (2010), 'dseriesname' refers to the first differenced form of the variable), there was no significant Granger causality of spectral centroid or dspectral centroid upon the corresponding change variable. On the other hand for arousal, there was causality from spectral centroid ($\chi^2(3) = 9.03$, $p < .029$) and residuals were satisfactory. This was also true for the stationarised (first differenced) model, ($\chi^2(2)$ 11.83, $p < .008$. More interestingly, given the relative lack of predictive power of the primary acoustic variables for perception of valence discussed in Dean and Bailes (2010), spectral centroid was Granger causal of valence ($\chi^2(3) = 11.03$, $p < .004$), and dspectral centroid of dvalence ($\chi^2(1) = 7.20$, $p < .007$).

To analyse this possible influence of spectral centroid further, we tested simultaneously the possible impact of both spectral flatness and spectral centroid as endogenous variables that might predict arousal or valence, as in previous multivariate assessments considered in our paper. For arousal, both spectral flatness ($\chi^2(4) = 35.39$ $p < .000$) and spectral centroid ($\chi^2(4) = 14.30$, $p < .006$) remained significant, and this was also true for the differenced model with respective p values <.000 and <.002 for flatness and centroid. There were similar results for valence and dvalence.

We next undertook these multivariate analyses with intensity as a third possible 'endogenous' predictor, given its pervasive and dominant influence in our previous analyses. While these multivariate models are generally detectably worse in BIC than those with 1 or 2 predictors, they are nevertheless informative of Granger causal interactions. For arousal, spectral centroid was again significant ($p < .013$) together with intensity ($p < .000$), while spectral flatness was not. Results were similar for modelling darousal (spectral centroid $p < .002$). In the case of valence, spectral centroid but not flatness remained Granger causal in this system tested with intensity. With valence itself, spectral centroid ($p < .001$) and intensity ($p < .024$) were causal. There were similar results for dvalence modelling.

Overall, these results suggested that spectral centroid may be a useful predictor in some circumstances in which spectral flatness is less so. Thus we assessed whether the addition of spectral centroid could enhance the core ARIMAX models of change, arousal and valence discussed within our paper (Dean & Bailes 2010).

Discriminating amongst New Candidate ARIMAX Models of Change, Arousal and Valence, evaluating Spectral Centroid as a Possible Predictor Variable

For perceived change in the music, as expected, spectral centroid could not enhance the models we described in Dean and Bailes (2010). The same was true for arousal, and when spectral centroid was included, the coefficients upon it were very small even though they were individually significant. Given the relative lack of good acoustic predictor variables for valence in the models described in Dean and Bailes (2010), with spectral flatness only modestly effective, spectral centroid was of particular interest here. We found that for valence the best available ARIMAX model was produced by dropping autoregressive lag 4 of dvalence and lag 12 of dspectral flatness from that in our paper, and adding lags 1 and 3 of dspectral centroid. This model had an AIC of 1055.7 and a BIC of 1118.6, thus showing improvement over the earlier model (AIC of 1062.6). The animate-sound impulse variable improved the ARIMAX model without spectral centroid, and was individually significant. However, the animate-sound impulse variable did not improve the ARIMAX model including spectral centroid, though remaining individually significant. These results suggest that spectral centroid may capture some of the features introduced by the animate-sound component which spectral flatness does not.

To consider further whether spectral centroid captures features which make significant contributions to these models we again assessed multivariate Vector Autoregressions in which the (undifferenced) variables are all treated as endogenous, and all acoustic and perceptual variables, including spectral centroid, are included (cf. Dean & Bailes, 2010, p. 167). Granger-causal parameters for perceived

change were as before (only arousal and intensity); whereas for arousal, valence and intensity were joined by spectral centroid as causal; and for valence, change and spectral centroid were causal but spectral flatness no longer so. The remaining question here is whether the Granger-causality of spectral centroid in the arousal and valence models is or is not associated with a significant impulse response: in other words, whether the influence is quantitatively significant given the other inputs. This was determined by the appropriate impulse response function analyses.

Impulse Response Function Analysis of the Impact of Spectral Centroid on Valence.

These interesting results, suggestive of an influence of spectral centroid on both arousal and valence, were assessed further by analysis of impulse response functions, treating all variables as endogenous, and using two autoregressive lags, which produces a highly significant overall model. However, Figure 1 shows such responses, and indicates that in spite of some additional Granger causalities, only autoregression (the impulse responses along the top-left to bottom-right diagonal), intensity and perceived change significantly influence other variables such that the response FEVD (fractional error variance decomposition) confidence limits cease to breach the zero line. As found earlier (Dean & Bailes, 2010, p. 167) intensity influences perceived change and arousal; while change influences valence. The other relationship displayed between intensity and spectral flatness is one between acoustic variables that are really exogenous to the experiment (that is, independent variables); and it has been discussed in some depth in Dean and Bailes (2010). At the most generous, one could interpret this relationship as suggesting that high intensity sounds are often constructed in this piece from high spectral flatness components. Results from the stationarised (first differenced) variables, again all entered into a VAR Impulse Response Functional Analysis are completely consistent with those using the native variables. We conclude that spectral centroid does not have significant predictive capacity for real-time arousal and valence perception in this piece. However, the positive results in some of the bivariate analyses discussed above show that it will be worthwhile in the future to consider spectral centroid along with spectral flatness.

Timbral Features of Music Encapsulated in Spectral Flatness and Spectral Centroid

The relative lack of success in using timbral variables other than perhaps the ecological features of human and animate sounds in these analyses of the Wishart piece, prompts a reconsideration of what is known of the perception of these features. Given space limitations, we can only make brief comments on this issue. It is fairly obvious that while the concept of 'pitch', can be roughly understood by most listeners, it is harder to grasp the pitch or perceptual centroid of an electroacoustic sound, and perceptual transparency is probably weaker still for the case of spectral flatness. Timbre is defined by the American National Standards Institute as the conglomerate of features which distinguish sounds which are identical in pitch and intensity. Thus we may ask to what extent perceptual centroid and flatness have perceptual relevance in complex music, and more particularly, with electroacoustic music involving sounds very different from those normally used in timbre discrimination studies. For example, the Wishart extract contains many noise components, comprises mostly inharmonic sounds, and lacks the sounds of musical instruments. Even its 'pitch' is probably often rather ambiguous. Before commenting on some recent literature on this we should make clear that in general we take the interaction between perception and cognition to be bidirectional: that is there can be perceptual processes which are primarily driven 'bottom-up', while others may be subject to much greater 'top-down' influences, which reflect prior experience and learning. This will not be elaborated here.

Literature on the perception of acoustic properties related to timbre is mainly based on responses to short individual tones (< 1 s in length) constructed largely of synthetic controlled mixtures of harmonic partials, or of the sounds of Western musical instruments. Commonly, a measurement of perceptual (dis)similarity between pairs of such sounds is made repeatedly, permitting the construction by multidimensional scaling of a distance map which best accommodates the combined data, and may itself be expressed in 2-4 dimensions, Euclidean or otherwise, so as to optimize the fit. As an excellent example, Caclin et al. (2005) used synthetic tones each having 20 harmonics, a duration 615 ms, and equated for pitch and perceptual loudness. Participants made dissimilarity ratings of tones varying in spectral centroid, spectral flux, spectral fine structure and attack time. Attack time, spectral centroid and fine structure emerged as 'major determinants'. It seems that spectral flatness would have been altered by both the last two factors. Another study revealed Garner interference between these three determinant factors, suggesting

crosstalk between the processing of multiple dimensions of timbre (Caclin et al., 2007). Few data seem to exist which deal with spectral flatness directly in this experimental context, in spite of its important position in the hierarchy of auditory classifiers in the MPEG-7 standard, which is in intention based on a perceptually optimizing approach. Comparatively few data deal with timbral properties during ecological music of minutes in duration or longer (though see Schubert (2004)). A notable exception is the study by analysis-by-synthesis of the influences of timbre on expressive clarinet performance (Barthet et al., 2011). This also deals with spectral centroid as the potentially primary timbral feature, and shows that removal of spectral centroid variations from pre-recorded performances resulted in 'the greatest loss of musical preference' (Barthet et al., 2011, p. 265). Again, however, the spectral centroid manipulations seem likely to have caused concomitant changes in spectral flatness. Perceptions of specific physical properties of timbre such as spectral centroid or flatness are clearly not entirely understood as yet.

Given this, we conclude that it remains important to seek to identify timbral measures which are useful to predict perceptions of musical structure and affect (and then perhaps to test empirically for their influence, as we have done in the case of intensity). As yet, we have not progressed far in this regard, and it might be important to continue to study more ecological concepts of timbre which describe features of sound source rather than acoustic properties of the sound (Bailes & Dean, in press). A broader issue arises from this: to what extent computational acoustic, or symbolic compositional features, or those extracted from music by statistical learning, are part of the perceptual-cognitive mediation chain that can translate sounding music into perceptual response, or merely analytical counterparts. We return to this issue in our final section.

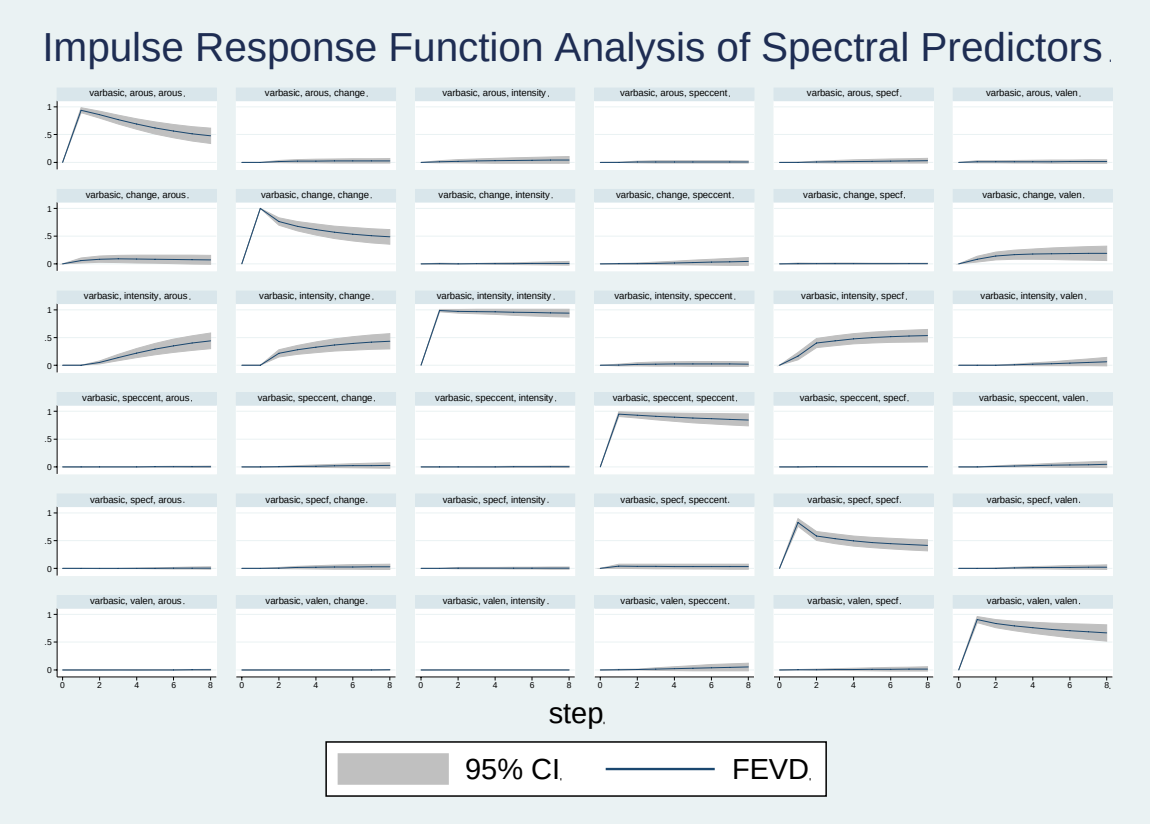

Fig. 1. Impulse Response Function Analysis of Spectral Predictors of the Grand Average Perceptual Time Series for Wishart's *Red Bird*. For each panel, 'varbasic' indicates the name given to the analysis, while the effect of unit change in the first named variable (the impulse) upon the second named (the response) is displayed over the next eight lags (i.e. four seconds). The shaded area represents the 95% confidence interval. Abbreviations: arous, arousal; speccent, spectral centroid; specf, spectral flatness; valen, valence.

COMPARISONS BETWEEN INDIVIDUAL AND GROUP-AVERAGE RESPONSES

In Dean and Bailes (2010) we chose to focus on grand average perceptual response time series, for reasons summarized there. However, we designed our participant groups to represent musicians with either generalist (M) or computer music/audio technology (EA, electroacoustic) musical skills, in comparison with a non-musician group (NM), bearing in mind the point subsequently raised by Pearce (2011), that group responses may hide subsets of perceptual strategies that differentiate individuals. Our groups and their individual members are compared in some detail in work in press (Bailes & Dean, in press) and in preparation. Here we illustrate these issues by a simple approach which is different from those we use in those papers. We study a randomly chosen individual from each expertise group in comparison with the grand average series, for influences on and between perceptual variables. This choice may allow the possibility of identifying distinct perceptual strategies adopted by different people. The approach we use here is to use the appropriate multivariate VAR/Impulse Response Function analyses, and to apply the same model as had been developed for the grand average perceptual series to each individual's series, to assess whether it retains significant explanatory power. Thus the VAR comprised change, arousal, valence, intensity, and two autoregressive lags, either using the grand average response series, or those from individuals EA1, M1, or NM1. The Impulse Response Functions are all qualitatively similar, and consistent with those of Figure 1. In some cases, the influence of intensity on change does not breach the zero value in terms of its confidence limits; in no cases are there any significant impulse responses not identified earlier. Table 1 shows that the key distinction between the individuals is whether perceived change is modelled well, as judged by the R-squared for its predictive equation; for a VAR, these R-squared parameters are an index of the degree to which the model fits the data for each particular response. This distinction in turn is predominantly a reflection of the extent of influence of intensity on perceived change as mentioned above, which is low for EA1 in comparison with the others and with the grand average, and successively higher for M1 and NM1. Thus, as we are investigating fully elsewhere, there are significant differences between individuals in their perceptual strategies. In general, we find in our larger study that the differences between individuals are rather greater than those between the groups, and tend to submerge inter-group differences. Putting this another way, in most respects inter-individual variation is comparably great in each expertise group, and represents the full spectrum of individual variation.

Table 1. Vector Autoregression analyses of individual and grand average perceptual time series

Equation		R^2	χ^2	$p > \chi^2$
Grand Average Responses	Arousal	0.9971	134349.5	0.0000
	Change	0.8093	1655.403	0.0000
	Valence	0.9978	174630.5	0.0000
EA1	Arousal	0.9785	17750.39	0.0000
	Change	0.0304	12.22054	0.1416
	Valence	0.9936	60094.54	0.0000
M1	Arousal	0.9947	72774.96	0.0000
	Change	0.5753	528.2246	0.0000
	Valence	0.9783	17607.08	0.0000
NM1	Arousal	0.9963	105508.2	0.0000
	Change	0.6838	839.1891	0.0000
	Valence	0.9921	48855.61	0.0000

Note. A set of 2-lag vector autoregressions was conducted, with perceived arousal, change and valence for either the grand average, EA1, M1 or NM1, and acoustic intensity treated as endogenous variables. There were nine parameters in each model (a constant together with two lags of each variable).

COMPUTATIONAL COGNITIVE MODELLING AND ITS POTENTIAL APPLICATION TO ELECTROACOUSTIC MUSIC

Johnson-Laird (1988) is one of the most forceful advocates of the view that a psychological process has not even been formulated, let alone understood, if one cannot express it in a precise computational format. In keeping with this tradition, Lewandowsky and Farrell (2011) in their recent stimulating book on computational modelling in cognition, illustrate with reference to mental rehearsal and Baddeley's working memory theory how computational modelling brings out the requirement for specifying many aspects which are left undefined in even recent verbal formulations of the concepts. As they say (Lewandowsky & Farrell, 2011, p. 25), an explanatory cognitive model should "describe all cognitive processes in great detail and leave nothing within their scope unspecified". Their 'scope' might for example be without regard for neural circuitry, though what they call 'cognitive architecture' models, such as ACT-R begin to address this circuitry.

As Pearce (2011) suggests, driving the IDyoM model (Pearce & Wiggins, 2006; Wiggins, Pearce, & Müllensiefen, 2009) to produce a minimised information content profile for a particular piece of symbolic music (such as their favoured note-centered minimal music, expressed in equal tempered notation) in a particular context of long term knowledge of a related corpus of music, may be to model how the brain statistically learns the nature of the piece and generates an expectation profile. That the IDyOM model can then successfully predict segmentation of the piece by certain listeners is, however, not a test of whether it is modelling the cognitive statistical learning process. Perhaps closer to such a test, and with positive results, is the recent study of Pearce et al. (2010), in which altering the information content of notes presented in a particular context did indeed produce correlated neural responses.

Returning to our own work, it is fairly clear that spectral centroid and spectral flatness bear a quite distant relationship to atomic perceptual processes, and it is still unclear how they may influence cognition. But acoustic intensity, on the other hand, is an immediate determinant of an important perceptual response, loudness, and this relationship is much better understood. Again, most studies use short tones, often synthetic, but it is clear that even with longer musical extracts, intensity is a close determinant of continuously perceived loudness. For example, we have recently presented evidence that the perception of loudness in a well-studied Dvorak Slavonic Dance is driven mainly 'bottom-up', bearing a very high correlation with intensity almost throughout the studied extract (Ferguson, Schubert, & Dean, 2011). Correspondingly we have been able to demonstrate by experimental manipulations of intensity profiles that in this particular piece, and in three other stylistic diverse pieces, intensity can be a major driver of perception of change and expressed arousal (Dean, Bailes, & Schubert, 2011). In the same paper we also showed that intensity might have actions in addition to those mediated via perceived loudness, a matter for future investigation. Our models of the influence of intensity upon perception of musical structure and affect are therefore candidate models of components of the cognitive processes involved in identifying musical change and affect. Contrary to Pearce's suggestion that our models are 'analytical' (a category he does not define fully), we would argue that they seek to be prototype cognitive models in much the same degree that information dynamic models do.

Bringing time series analysis, acoustic parameters, performance parameters, and information dynamics together in future work will be a useful step, and one in which we are engaged already. In other work in preparation on an unmeasured Prelude for Harpsichord by Couperin, involving Pearce and Wiggins and initiated by Gingras and collaborators in Canada, we have already obtained evidence as to the power of both information content and entropy based on pitch structure in predicting performance timing, and the subsequent impact of timing parameters on perceptions of tension. Roles for intensity in this system are yet to be addressed, but the harpsichord is an instrument with an unusually narrow dynamic range. Developing the assessment of information rate in electroacoustic music from the contributions of Dubnov and others mentioned by Pearce (2011) will be a complementary challenge, and will permit us to undertake information dynamic studies with time series analysis in relation to our target pieces.

REFERENCES

Bailes, F., & Dean, R. T. (in press). Comparative time series analysis of perceptual responses to electroacoustic music. *Music Perception.*

Barthet, M., Depalle, P., Kronland-Martinet, R., & Ystad, S. (2011). Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance. *Music Perception*, Vol. 28, No. 3, pp. 265-278.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, Vol. 118, No. 1, pp. 471-482.

Caclin, A., Giard, M.-H., Smith, B. K., & McAdams, S. (2007). Interactive processing of timbre dimensions: A Garner interference study. *Brain Research*, Vol. 1138, pp. 159-170.

Dean, R. T., & Bailes, F. (2010). Time series analysis as a method to examine acoustical influences on real-time perception in music. *Empirical Musicology Review*, Vol. 5, No. 4, pp. 152-175.

Dean, R. T., Bailes, F., & Schubert, E. (2011). Acoustic intensity causes perceived changes in arousal levels in music. *PLoS One*, Vol. 6, No. 4.

Ferguson, S., Schubert, E., & Dean, R. T. (2011). Continuous subjective loudness responses to reversals and inversions of a sound recording of an orchestral excerpt. *Musicae Scientiae*. doi: 10.1177/1029864911410122

Johnson-Laird, P. N. (1988). *The Computer and the Mind: An Introduction to Cognitive Science*. London: Fontana Press.

Landy, L. (2009). Sound-based music 4 all. In: R. T. Dean (Ed.), *The Oxford Handbook of Computer Music*. New York: Oxford University Press, pp. 518-535.

Lewandowsky, S., & Farrell, S. (2011). *Computational Modeling in Cognition: Principles and Practice*. Los Angeles, London, New Delhi, Singapore, Washingtone DC: Sage.

Pearce, M. T. (2011). Time-series analysis of music: Perceptual and information dynamics. *Empirical Musicology Review*. Vol. 6, No. 2, pp. 125-130.

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, Vol. 50, No. 1, pp. 302-313.

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, Vol. 23, No. 5, pp. 377-405.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception*, Vol. 21, No. 4, pp. 561-585.

Wiggins, G. A., Pearce, M. T., & Müllensiefen, D. (2009). Computational modeling of music cognition and musical creativity. In: R. T. Dean (Ed.), *The Oxford Handbook of Computer Music*. New York: Oxford University Press, pp. 383-420.