Connection Science
Vol. 00, No. 00, January 2008, 1–32

RESEARCH ARTICLE

Information Dynamics: Patterns of expectation and surprise in the perception of music

Samer Abdallah and Mark Plumbley
Centre for Digital Music, Queen Mary, University of London,
Mile End Road, London E1 4NS
(*Received 00 Month 200x; final version received 00 Month 200x*)

Measures such as entropy and mutual information can be used to characterise random processes. In this paper, we propose the use of several *time-varying* information measures, computed in the context of a probabilistic model which evolves as a sample of the process unfolds, as a way to characterise temporal structure in music. One such measure is a novel *predictive information rate* which we conjecture may provide a conceptually simple explanation for the 'inverted-U' relationship often found between simple measures of randomness (e.g. entropy rate) and judgements of aesthetic value (1). We explore these ideas in the context of Markov chains using both artificially generated sequences and two pieces of minimalist music by Philip Glass, showing that even such a manifestly simplistic model (the Markov chain), when interpreted according to information dynamic principles, produces a structural analysis which largely agrees with that of an expert human listener. Thus, we propose that our approach could form the basis of a theoretically coherent yet computationally plausible model of human perception of formal structure, potentially including seemingly abstract qualities like interestingness and aesthetic goodness.

Keywords: information theory; expectation; surprise; subjective probability; Bayesian inference; Markov chain; music.

1. Expectation and surprise in music

One of the more salient effects of listening to music is to create *expectations* of what is to come next, which may be fulfilled immediately, after some delay, or not at all as the case may be. This is the thesis put forward by, amongst others, music theorists L. B. Meyer (2) and Narmour (3). In fact, this insight predates Meyer quite considerably; for example, it was elegantly put by Hanslick (4) in the nineteenth century:

> 'The most important factor in the mental process which accompanies the act of listening to music, and which converts it to a source of pleasure, is frequently overlooked. We here refer to the intellectual satisfaction which the listener derives from continually following and anticipating the composer's intentions—now, to see his expectations fulfilled, and now, to find himself agreeably mistaken. It is a matter of course that this intellectual flux and reflux, this perpetual giving and receiving takes place unconsciously, and with the rapidity of lightning-flashes.'

An essential aspect of this is that music is experienced as a phenomenon that 'unfolds' in time, rather than being apprehended as a static object presented in its entirety. Meyer argued that musical experience depends on how we change and revise our conceptions *as events happen*, on how expectation and prediction interact

Corresponding author; email: samer.abdallah@elec.qmul.ac.uk

ISSN: 0954-0091 print/ISSN 1360-0494 online

DOI: 10.1080/09540090xxxxxxxxxxxx
http://www.informaworld.com

with occurrence, and that, to a large degree, the way to understand the effect of music is to focus on this 'kinetics' of expectation and surprise.

The business of making predictions and assessing surprise is essentially one of reasoning under conditions of uncertainty and manipulating degrees of belief about the various proposition which may or may not hold, and, as has been argued elsewhere (5, 6), best quantified in terms of Bayesian probability theory. Thus, we suppose that when we listen to music, expectations are created on the basis of our familiarity with various stylistic norms encode the statistics of music in general, the particular styles of music that seem best to fit the piece we happen to be listening to, and the emerging structures peculiar to the current piece. There is experimental evidence that human listeners are able to internalise statistical knowledge about musical structure, e.g. (7, 8), and also that statistical models can form an effective basis for computational analysis of music, e.g. (9–11).

1.1. *Music and information theory*

Given a probabilistic framework for music modelling and prediction, it is a small step to apply quantitative information theory (12) to the models at hand. The relationship between information theory and music and art in general has been the subject of some interest since the 1950s (2, 13–17). The general thesis is that perceptible qualities and subjective states like uncertainty, surprise, complexity, tension, and interestingness are closely related to information-theoretic quantities like entropy, relative entropy, and mutual information. Berlyne (1) called such quantities 'collative variables', since they are to do with patterns of occurrence rather than medium-specific details, and developed the ideas of 'information aesthetics' in an experimental setting.

Previous work in this area (18) treated the various information theoretic quantities such as entropy as if they were intrinsic properties of the stimulus—subjects were presented with a sequence of tones with 'high entropy', or a visual pattern with 'low entropy'. These values were determined from some known 'objective' probability model of the stimuli,[1]or from simple statistical analyses such as computing emprical distributions. Our approach is explicitly to consider the role of the observer in perception, and more specifically, to consider estimates of entropy etc. with respect to *subjective* probabilities.

More recent work on using information theoretic concepts to analyse music in includes Simon's (20) assessments of the entropy of Jazz improvisations and Dubnov's (21–23) investigations of the 'information rate' of musical processes, which is related to the notion of redundancy in a communications channel. Dubnov's work in particular is informed by similar concerns to our own and we will discuss the relationship between it and our work at several points later in this paper (see § 2.6, § 5.1 and § 6).

1.2. *Information dynamic approach*

Bringing the various strands together, our working hypothesis is that as a listener (to which will refer gender neutrally as 'it') listens to a piece of music, it maintains a dynamically evolving statistical model that enables it to make predictions about how the piece will continue, relying on both its previous experience of music and the immediate context of the piece. As events unfold, it revises its model and hence

[1]The notion of objective probabalities and whether or not they can usefully be said to exist is the subject of some debate, with advocates of subjective probabilities including de Finetti (19). Accordingly, we will treat the concept of a 'true' or 'objective' probability models with a grain of salt and not rely on them in our theoretical development.

its probabilistic belief state, which includes predictive distributions over future observations. These distributions and changes in distributions can be characterised in terms of a handful of information theoretic-measures such as entropy and relative entropy. By tracing the evolution of a these measures, we obtain a representation which captures much of the significant structure of the music. This approach has a number of features which we list below.

(1) *Abstraction*: Because it is sensitive mainly to *patterns* of occurence, rather the details of which specific things occur, it operates at a level of abstraction removed from the details of the sensory experience and the medium through which it was received, suggesting that the same approach could, in principle, be used to analyse and compare information flow in different temporal media regardless of whether they are auditory, visual or otherwise.

(2) *Generality*: This approach does not proscribe which probabilistic models should be used—the choice can be guided by standard model selection criteria such as Bayes factors (24), etc.

(3) *Richness*: It may be effective to use a model with time-dependent latent variables, such as a hidden Markov model. In these cases, we can track changes in beliefs about the hidden variables as well as the observed ones, adding another layer of richness to the description while maintaining the same level of abstraction. For example, harmony (i.e., the 'current chord') in music is not stated explicitly, but rather must be inferred from the musical surface; nonetheless, a sense of harmonic progression is an important aspect of many styles of music.

(4) *Subjectivity*: Since the analysis is dependent on the probability model the observer brings to the problem, which may depend on prior experience or other factors, and which may change over time, inter-subject variablity and variation in subjects' responses over time are fundamental to the theory. It is essentially a theory of subjective response

Having outlined the basic ideas, our aims in pursuing this line of thought are threefold: firstly, to propose dynamic information-based measures which are coherent from a theoretical point of view and consistent with the general principles of probabilistic inference, with possible applications in regulating machine learning systems; secondly, to construct computational models of what human brains are doing in response to music, on the basis that our brains implement, or at least approximate, optimal probabilistic inference under the relevant constraints; and thirdly, to construct a computational model of a certain restricted field of aesthetic judgements (namely judgements related to formal structure) that may shed light on what makes a stimulus interesting or aesthetically pleasing. This would be of particular relevance to understanding and modelling the creative process, which often alternates between generative and selective or evaluative phases (25), and would have applications in tools for computer aided composition.

1.3. *Outline of the paper*

The remainder of the paper is organised as follows: in § 2 we provide general definitions of the information measures that we are going to examine; in § 3 we show how these measures can be computed for a particular model, the Markov chain, and examine the information dynamics of sequences generated artificially from known Markov chains; in § 4 we examine the information dynamics of online learning in Markov chains. We apply the Markov chain model to minimalist music by Philip Glass § 5, and show how the information-dynamic approach yields a plausible structural analysis that largely agrees with that of a human expert listener. In § 6 we are in a position to discuss our approach in relation with previous work in the same area. We wrap-up with conclusions and future directions in § 7.

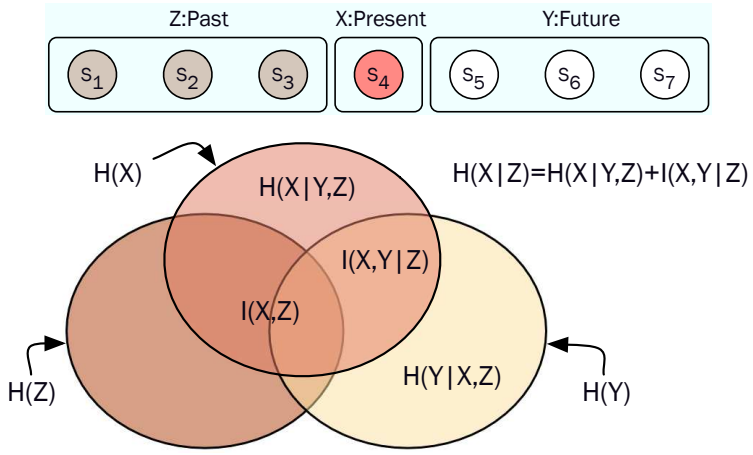

Figure 1. By grouping the elements of a random sequence into a *past*, *present*, and *future*, we can consider a number of information measures, some of which are well known, like the entropy rate $H(X|Z)$, and some of which have not, to our knowledge, been investigated before, such as the average predictive information rate $I(X, Y|Z)$. The relationship between several such measures can be visualised as a Venn diagram. Note that Z and Y actually stand for the *infinite* past and future and are only shown as finite for visualisation purposes.

2. Model-based observation of random processes

In this section we define some of the information measures that a model-based observer can compute given a realisation of a random process and a statistical model that can be updated dynamically as the process unfolds. This observer-centric view highlights the point that the probabilities we consider here are essentially *subjective* probabilities, and do not require any 'objective' or frequentist interpretation. The observer's model need not be the 'correct' one, and we need not rely on the epistomologically questionable notion of a 'correct' model existing (6).

Consider a snapshot of a stationary random process taken at a certain time: we can divide the timeline into an infinite 'past' and 'future', and a notional 'present' of finite duration. Observations of the process can be grouped into three random variables, say Z, Y, and X, corresponding to these three time intervals respectively (see fig. 1). The model is summarised by the observer's probability distribution $p_{XY|Z}$ over the present and future given the past. For discrete variables, $p_{XY|Z}(x, y|z)$ is the probability with which the observer expects to see x followed by y given that it has already seen z. We can now consider how the observer's belief state evolves when it learns that $X{=}x$.

2.1. *'Surprise'-based measures*

To obtain a first set of four information measures, we marginalise out the future Y to get the distribution for the immediate prediction, $p_{X|Z}$. The negative log-probability

$$\mathcal{L}(x|z) \triangleq -\log p_{X|Z}(x|z), \tag{1}$$

can be thought of as the 'surprisingness' of x in the context of z. The expectation of this quantity (given a particular z) is the entropy of the predictive distribution, which we will write as $H(X|Z{=}z)$ to emphasise that it is a function of the observed past z; it is a measure of the observer's uncertainty about X before the observation is made, and quantifies the notion that certain contexts z may lead the observer to 'expect the unexpected'.

Once the observer sees that $X{=}x$, it can compute its surprisingess $\mathcal{L}(x|z)$, but for some classes of model it may be possible to average $\mathcal{L}(x|z)$ over the past contexts (given the current model) that could have lead to the current observation, that is, over $Z|X{=}x$. This average in-context surprisingness of the symbol x might be useful as a sort of static analysis of the model, helping to pick out which are the most significant states in the state space. By averaging $\mathcal{L}(x|z)$ over *both* variables, we obtain the conditional entropy $H(X|Z)$, which, bearing in mind that Z stands for the *infinite* past, is equivalent to the entropy rate of the process according to the observer's current model. Thus, the first four measures are the surprisingness and its three averages over $(X|Z{=}z)$, $(Z|X{=}x)$, and (X, Z) jointly.

2.2. *Predictive information-based measures*

Perhaps more important than intrinsic surprisingness of an observation is the information it carries *about* about the unobserved future, *given* that we already know the past. This is what we are calling the *predictive information* (PI). Hence, to obtain a second set of four information measures, we consider the information supplied about Y by the observation that $X{=}x$, given that we already know $Z{=}z$, quantified as the Kullback-Leibler (KL) divergence between the predictive distribution over Y before and after the event $X{=}x$, that is,

$$\mathcal{I}(x|z) \triangleq I(X{=}x, Y|Z{=}z) = D(p_{Y|X=x,Z=z}||p_{Y|Z=z}), \tag{2}$$

where $p_{Y|Z=z}(y) = \int p_{XY|Z=z}(x, y)\ \mathrm{d}x$ and $D(\cdot||\cdot)$ is the KL divergence between two distributions[1]. Like $\mathcal{L}(x|z)$, this is a function of the observations z and x, and we can take expectations over X or Z or both. Averaging over the prediction $X|Z{=}z$, that is, computing $\mathrm{E}_{X|Z=z}\mathcal{I}(X|z)$, tells us the amount of new information we *expect* to receive from the next observation about the future. It could be useful as a guide to how much attention needs to be directed towards the next event even before it happens. This is different from Itti and Baldi's proposal that Bayesian *surprise* attracts attention (27), as it is a mechanism which can operate *before* the surprise occurs.

The average of the PI over preceeding contexts $Z|X{=}x$, that is, the expectation $\mathrm{E}_{Z|X=x}\mathcal{I}(x|Z)$, is the amount of information about the future carried, on average, by each value in the state space of X. As before, this tells us something about the significance of each symbol in the alphabet, picking out which symbols tend to be most informative about the future. One might predict that these states will tend to appear as 'onset' states, or as the 'foreground' against a 'background' of the states that tend not to carry much information.

Averaging over both X and Z gives us the *predictive information rate* (PIR), which is, for a given random process model, the average rate at which new information arrives about the future. The expression reduces to what one might call a 'conditional mutual information' (see fig. 1):

$$I(X, Y|Z) = H(Y|Z) - H(Y|X, Z). \tag{3}$$

Since Y represents the infinite future, we would, in general, expect both entropy terms on the right to diverge, and so it is preferable to define the PIR using either

[1]Note that here and elsewhere in the paper, we are using integrals in generalised (Lebesgue) sense to cover both continuous and discrete random variables. The relevant probability density functions are defined in terms of the Radon-Nikodym derivative and in the discrete case reduce to the familiar discrete distribution (26, p. 25).

of two equivalent forms:

$$\begin{aligned} I(X,Y|Z) &= H(X|Z) - H(X|Y,Z) \\ &= I(\{Z,X\},Y) - I(Z,Y), \end{aligned} \tag{4}$$

where $I(\{Z,X\},Y)$ is the mutual information between Y and the *pair* of variables $\{Z,X\}$. Note the difference between the predictive information defined in (2), which is a function of past and present observations (and the observer's model), and the predictive information rate defined in (4), which is the *average* of the predictive information with respect to the observer's model and is a function of that model only.

Overall, the four measures in the second set are $\mathcal{I}(x|z)$ and its expectations over $(X|Z{=}z)$, $(Z|X{=}x)$, and (X,Z) jointly. Unlike those in the first set, these measures are computed in terms of KL divergences and hence are invariant to invertible transformations of the observation spaces: for continuous random variables, the random process could be 'transcoded' using different symbols and perhaps a different modality, and as long as the transcoding was invertible, the predictive information measures would remain the same. Even for discrete random variables, the uniqueness of the entropy as an absolute and invariant measure of uncertainty can be questioned. The standard definition of the entropy implicitly assumes that the uniform distribution is the 'most uncertain' and therefore embodies the least prior information, but as Bernardo and Smith point out (28, p. 79), the choice of the uniform distribution as a reference measure, though seemingly natural, should not be automatic. For example, consider a discrete space of possibilities $\{a,b,c\}$ where b and c are somehow semantically similar—perhaps they are variants of a broader class $\{b,c\}$. In this case, a state of ignorance might be better represented by making the two broad classes equiprobable, with $p(a) = \frac{1}{2}$ and $p(b) = p(c) = \frac{1}{4}$. Thus, information measures based on entropy, rather than mutual information and KL divergences, necessarily have a certain amount of arbitrariness built into them due to implicit choices about the representation and the implicit reference measure for the event space.

2.3. *Information about model parameters*

Finally, another information measure can be obtained by considering an observer using an explicitly parameterised model. In this case, the observer's belief state would include a probability distribution for the parameters Θ. Each observation would cause a revision of that belief state and hence supply information about the parameters, which we will again quantify as the KL divergence between prior and posterior distributions $D(p_{\Theta|X=x,Z=z}||p_{\Theta|Z=z})$. We call this the 'model information rate'.

Note that in a rigorous analysis of the predictive information in a model which includes unknown parameters, information gained about the parameters would also manifest itself as information gained about future observations, since the correct way to compute the probability of future observations in these models is to take account of uncertainty about the parameters and integrate them out. This can be done for certain models where there is a fixed set of static parameters and the observations are conditionally independent given the parameters (29), but in most cases an exact computation will be intractible.

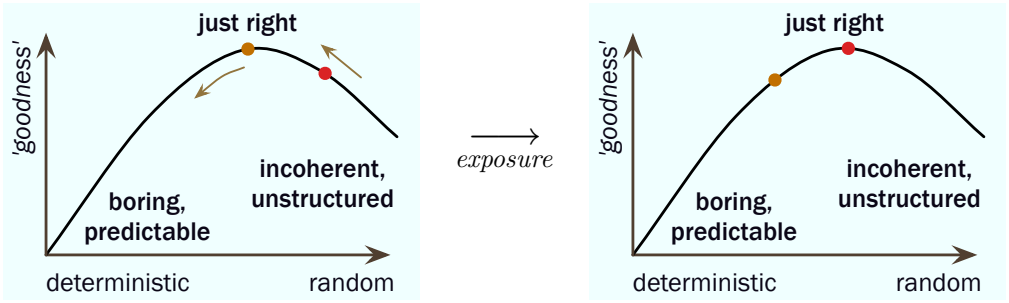

Figure 2. Relationship between apparent complexity and aesthetic value, and the change in value judgement which sometimes occurs after repeated exposure.

2.4. *Predictive information rate as a measure of structure*

Many studies looking into the relationship between stochastic complexity as measured by entropy or entropy rate, and what has variously been called 'pleasingness', 'hedonic values', and 'aesthetic value', reveal an inverted 'U' shaped curve (see fig. 2) where the highest value is attached to processes of intermediate entropy (1).

This type of relationship (though not in quantitative information-theoretic terms) was also observed by Wundt (30). Intuitively, patterns which are too deterministic and ordered are boring, while those which are too random are perceived as unstructured, featureless, and, in a sense, 'uniform', in the way that white noise is. Hence, a sequence can be uninteresting in two opposite ways: by being utterly predictable *or* by being utterly unpredictable. Meyer (31) hints at the same thing while discussing the relation between the rate of information flow and aesthetic experience, suggesting that 'If the amount of information [by which he means entropy and surprisingness] is inordinately increased, the result is a kind of cognitive white noise.'

The explanations for this usually appeal to a need for a 'balance' between order and chaos, unity and diversity, and so on, in a generally imprecise way. However, the predictive information rate (4) seems to incorporate this balance automatically (see fig. 3), acheiving a maximum for sequences which are neither deterministic nor totally uncorrelated across time. Our interpretation of this is that when each event appears to carry no new information about the unknown future, it is not worth attending to, and in a way, meaningless. More precisely, such events are *useless* for the business of dealing with the future.

Indeed, the idea of using utility to determine information processing has been examined by previous researchers. For example, Bernardo and Smith (28, p. 79) *derive* Bayesian inferance from a more general utility theory by treating inference as a decision problem, which is to choose an inferential distribution from the class available given that each choice has a score or utility if a particular state of affairs turns out to be true. In this theory, the KL divergence, which we have been using as a measure of information gain, becomes equivalent to an expected gain in utility.

In a practical setting, Levitt *et al* (32) use utility theory in an active sensing system to help choose which observation gathering actions to perform next. In a system with limited computational or sensing capabilities, estimation of the amount of information to be gained from one of several possible observations can help choose what is likely to be the most profitable course of action. This is very reminiscent of Gibson's theory of 'active perception' (33), wherein a perceptual system does not merely receive sensory data in a passive way, but actively directs the sensory apparatus to seek out the most promising regions of the sensory field. An obvious example is the way we direct our visual attention by looking *at* objects of interest, placing the image of the object on the fovea where visual acuity is greatest.

Returning to the subject of perceived qualities of sequences, Berlyne (1, ch. 13) also discusses the effect which repeated exposure to a stimulus has on its perceived aesthetic value. The evidence he reviews is conflicting: in some cases repeated exposure leads to an increase in preference, while in others, to a decrease. Berlyne argues that this can be understood as a process of migration to the left along the Wundt curve as the subjective complexity of the stimuli decreases due to the observer learning something of its structure, as shown in fig. 2. Stimuli starting out on the right of the curve will be liked more as the observer becomes more familiar with them, while those starting near the top or on the left will be liked less and less. We will return to this in § 4 where we show that the predictive information rate can display similar behaviour when computed using a probability model that adapts over time.

2.5. *The status of models, parameters, and observations*

In the above definitions of the various information measures, we have made use of the notion of *models*, possibly parameterised, as distinct from observations, and in particular, as distinct from the aggregate of past observations to which we have been referring as z, that is, the *particular* observered values of the random variable Z. This is a good point to discuss the ontological status and validity of such a distinction. Some of the measures, namely the surprisingness $\mathcal{L}(x|z)$ (1), the predictive information $\mathcal{I}(x|z)$ (2), and their respective expectations with respect to the observer's predictive distribution for the next observation $(X|Z{=}z)$, are functions of the particular history of the observer as represented by the proposition $Z{=}z$ and the observer's subsequent expectations about X and the future Y. The observer's model is only implicated in so far as it is the mechanism by which an expectation is generated; its internal details are not important. In contrast, the other measures involve drawing a line between observed data and any other variables used to represent a parameterised model, which are then thought of as existing in some sense independently of the data. Once this is done, we can average, in an essentially counterfactual way, over *hypothetical* histories while keeping the model constant, to obtain the entropy rate $H(X|Z)$ and the predictive information rate $I(X,Y|Z)$, which are properties of the model and not of any particular observations. In addition, we can also consider changes of the model; this is how we obtain the model information rate, since it is a function of two consecutive sets of beliefs about (i.e. distributions over) the model parameters.

The problem is that this division between data and model parameters is somewhat indistinct: the model is but a summary of past data; if we erase the observations but keep the model parameters, we retain some, and possibly much, information about the observations. In a modular cognitive system, most of what is presented as 'data' at one level, including abstract or symbolic representations of stimuli, is the result of processing in a previous level and can be considered as 'model' or 'parameters' in that previous level. Thus, we would argue that measures based on manipulations of a model independently from the observations that were used to construct it are not quite on the same level of generality and abstraction as the measures that do not explicitly invoke a model, but only specific expectations based on specific observations. The model information rate, for example, is suggested only as an easily computed proxy for a more complete analysis of how information gained about model parameters would manifest itself as information gained about future observations (i.e. predicitive information). For example, the fact that the model information rate gives such good in results in the analysis of *Two Pages* (to come in § 5.2.1) indicates the potential benefit of a more thorough analysis of predictive information in non-stationary Markov chains.

2.6. *Redundancy and Dubnov's 'Information rate'*

Dubnov (21) defines the 'information rate' (IR) as the mutual information between the past and the present, or in our notation, $I(X, Z) = H(X) - H(X|Z)$, which is essentially a measure of the redundancy in the sequence, as discussed by Attneave (34). It is the difference between the actual entropy rate and the entropy rate that would be achieved if all temporal dependencies were removed, and quantifies a dependency between parts. According to this definition, X and Z are to be considered random variables whos values are unknown when computing the IR, but have a certain assumed joint distribution. That is, the IR is the expected information to be gained from Z (the 'past') about X (the 'present') evaluated *before* any observations are made and in terms of the observer's prior beliefs expressed by the probability distribution $p(z, x)$. Its value is therefore a function of the observer's prior state of mind only and does not express any property of the actually observed data. It also requires that a certain moment in time be selected as the beginning of 'the past', so that the observer's prior belief state $p(x, z)$ can be identified.

In a recent paper (23), Dubnov presents a method for computing the IR of a process in terms of two components, the 'data-IR' and the 'model-IR'. The data-IR was introduced previously (21, 22) and involves fitting a sequence of models to successive segments of the input and computing the redundancy of these models. Dubnov (23) notes that this ignores the role of the data in learning the model for each segment and therefore introduces the model-IR to account for information in the data about the model parameters which are being learned from that data. However, since the model-IR is computed for 'macro-blocks' of audio data approximately 5 s long, 'the past' Z is not the entire history of the signal, but only the last 5 s, a period which Dubnov suggests is representative of the 'perceptual present'. Observations made before the current macro-block, which we can refer to as $W{=}w$, provide the context for the model which is about to be fitted to the current macro-block. With this dependency made explicit, the IR is $I(Z, X|W{=}w)$, which we can now identify as the *expected* predictive information in the 'perceptual present' Z about a very short piece of the future X given the actually observed past $W{=}w$.

Dubnov's method for computing the model-IR is difficult to evaluate as there appears to be an error in the derivation and it involves some questionable or inconsistent assumptions (35). However, at the end of the process and having introduced the model-IR as a correction to data-IR, Dubnov does not sum the two and examine the properties of this improved approximation to the IR, but rather continues to treat them separately, even suggesting that the *difference* between the two approximations is a significant quantity that indicates perceived interestingness.

We will discuss Dubnov's information rate further in § 6; in particular we will examine whether or not the IR exhibits the claimed 'inverted-U' behaviour. Except when we are discussing Dubnov's work directly, we will refer to $I(X, Z)$ as the 'redundancy' rather than the 'information rate', both to avoid confusion with our predictive information rate and because, we would argue, it is not meaningful to think of it as a rate of arrival or accumulation of information (in the sense of reduction of uncertainty as discussed in § 2.2) about any particular thing.

2.7. *Role of multiple information measures*

Thus far, we have defined 9 different information measures (10 if we include the redundancy), and one may ask why we need so many. Part of the answer is that we are not yet sure which of these, if any, will prove to be psychologically or practically relevant. However, a case for investigating multiple measures can be made on more principled grounds as well for pragmatic reasons. Firstly, it would be unreasonable

to assume that the human response to music is one dimensional, and it is important to determine whether or not we use words such as 'interestingness', 'predictabilitiy', 'complexity' etc. to describe distinct qualities that are nonetheless 'collative' and susceptible to the kind of statistical analysis we are advocating. Secondly, the different measures may have distinct rôles in implementing self-regulation mechanisms in practical machine learning systems. Thirdly, providing a multidimensional characterisation of the stimulus all at the same conceptual level makes room for a greater variety of patterns to emerge during the processing of a given stimulus. Even if the individual dimensions are not interpreted, this makes room for a rich language of informational 'gestures' that can be composed into a stimulus and recognised by the observer.

3. Information dynamics in Markov chains

To illustrate the how the measures defined in § 2 can be computed in practice, we will consider one of the simplest random processes, a first order Markov chain. Let S be a Markov chain with a finite state space $\{1, \ldots, N\}$ such that S_t is the random variable representing the t^{th} element of the sequence. The model is parameterised by a transition matrix $a \in \mathbb{R}^{N\times N}$ encoding the distribution of any element of the sequence given previous one, that is $p(S_{t+1}{=}i|S_t{=}j) = a_{ij}$. Since we require the process to be stationary, we set the distribution for the initial element S_1 to the equilibrium distribution of the transition matrix, that is, $p(S_1{=}i) = \pi^a_i$ where π^a is a column vector satisfying $a\pi^a = \pi^a$. To ensure that the equilibrim distribution is unique, we also require that the Markov chain be ergodic. Under these conditions, the Markov chain will have an entropy rate which can be written as a function of a alone:

$$\dot{\mathcal{H}} : \mathbb{R}^{N\times N} \to \mathbb{R}, \qquad \dot{\mathcal{H}}(a) = \sum_{i=1}^{N} \pi^a_i \sum_{j=1}^{N} -a_{ji} \log a_{ji}. \tag{5}$$

The assumptions of stationarity and ergodicity are made so that the resulting entropy and predictive information rates are characteristic of the process as a whole rather than specific to particular times during the evolution of the Markov chain, and also so that these average rates are in principle accessible to an observer exposed to a sufficiently long sample of the process. A non-ergodic Markov chain may have an initial transient phase and can get stuck in part of the state space, in which case the observer will not get a representative view of the process no matter how long the sample.

The Markov dependency structure means that, for the purposes of computing the measures defined in § 2, the 'past' and 'future' at time t can be collapsed down to the previous and next elements of the chain (see appendix). In terms of our earlier notation, we can set $Z = S_{t-1}$, $X = S_t$, and $Y = S_{t+1}$. Equations (7) and (8) below give expressions for the eight information measures from the first two sets defined in § 2. Some of these are expressed in terms of the 'time-reversed' transition matrix defined as

$$a^\dagger_{ij} = p(S_{t-1}{=}j|S_t{=}i) = a_{ij}\pi^a_j/\pi^a_i. \tag{6}$$

Note that the over- and under-bars are intended as mnemonics for the expectations over S_t and S_{t-1} respectively.

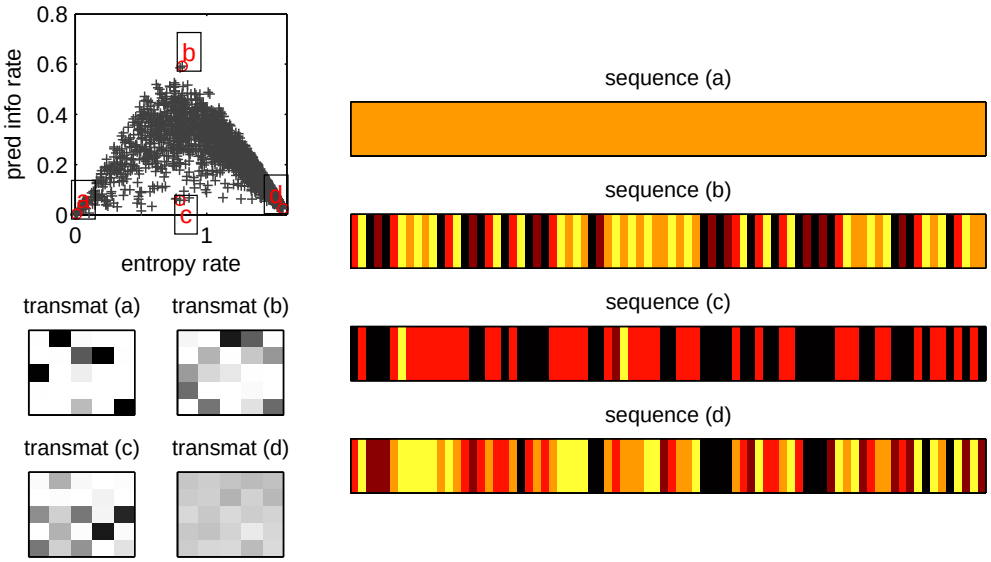

Figure 3. The space of transition matrices explored by generating them at random and plotting entropy rate vs PIR. (Note inverted 'U' relationship). Four of the transition matrices are shown along with sample sequences. Sequence (a) is simply the endless repetition of state 4. Matrix (d) is almost uniform. Matrix (b) has the highest PIR.

The first four, 'surprise'-based, measures are

$$\begin{array}{lll} \mathcal{L}(i|j) & = -\log p(S_t{=}j|S_{t-1}{=}i) & = -\log a_{ij}, \\ \overline{\mathcal{L}}(j) & = \mathrm{E}_{\ i\sim S_t|S_{t-1}=j}\mathcal{L}(i|j) & = \sum_{i=1}^{N} a_{ij}\mathcal{L}(i|j), \\ \underline{\mathcal{L}}(i) & = \mathrm{E}_{\ j\sim S_{t-1}|S_t=i}\mathcal{L}(i|j) & = \sum_{j=1}^{N} a^{\dagger}_{ij}\mathcal{L}(i|j), \\ \overline{\underline{\mathcal{L}}} & = H(S_{t+1}|S_t) & = \dot{\mathcal{H}}(a). \end{array} \tag{7}$$

The second four, predictive-information-based, measures are

$$\begin{array}{lll} \mathcal{I}(i|j) & = D(p_{S_{t+1}|S_t=i}||p_{S_{t+1}|S_{t-1}=j}) & = \sum_{k=1}^{N} a_{ki}(\log a_{ki} - \log[a^2]_{kj}), \\ \overline{\mathcal{I}}(j) & = \mathrm{E}_{\ i\sim S_t|S_{t-1}=j}\mathcal{I}(i|j) & = \sum_{i=1}^{N} a_{ij}\mathcal{I}(i|j), \\ \underline{\mathcal{I}}(i) & = \mathrm{E}_{\ j\sim S_{t-1}|S_t=i}\mathcal{I}(i|j) & = \sum_{j=1}^{N} a^{\dagger}_{ij}\mathcal{I}(i|j), \\ \overline{\underline{\mathcal{I}}} & = I(S_t, S_{t+1}|S_{t-1}) & = \dot{\mathcal{H}}(a^2) - \dot{\mathcal{H}}(a). \end{array} \tag{8}$$

Note that the result that the PIR is the difference between the entropy rates of the one- and two-step Markov chains does not generalise to non-Markovian dynamics. It is due to the fact that in a Markov chain, any information gained about the infinite future is entirely accounted for by the information gained about the next single element of the chain.

An example of an analyis of a Markov chain (both the transition matrix and a sampled state sequence) using the surprise and predicitive information based measures is shown in fig. 4. It shows how surprisingness and predictive information are distinct quantities (e.g. at time points 40, 49, 65 and 80), and also how the five states have distinct statistical properties, occupying distinctive positions in the per-state analysis shown in the lower three panels.

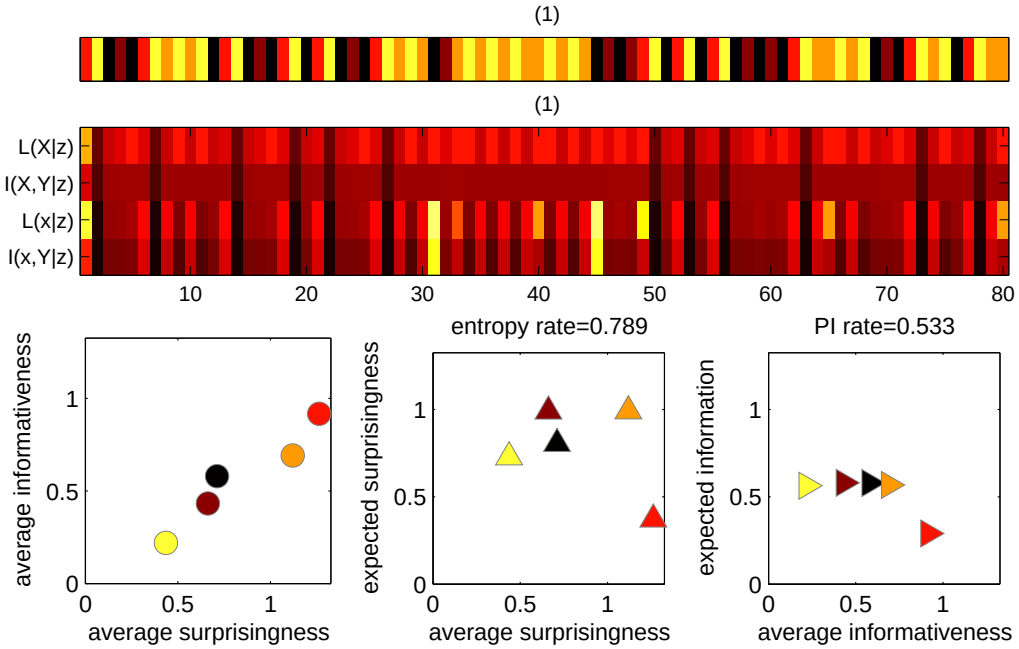

Figure 4. Analysis of transition matrix and sequence (b) from fig. 3. The upper panels show the sequence itself along with the dynamic evolution of, from top to bottom, the predictive uncertainty $\overline{\mathcal{L}}(j)$, the expected predictive information $\overline{\mathcal{I}}(j)$, the surprisingness $\mathcal{L}(i|j)$ and the predictive information $\mathcal{I}(i|j)$. The lower panels summarise the static analysis of the system, plotting the average suprisingness $\underline{\mathcal{L}}(i)$, the average informativeness $\underline{\mathcal{I}}(i)$, the predictive uncertainty $\overline{\mathcal{L}}(i)$, and the information expectancy $\overline{\mathcal{I}}(i)$ generated by each of the five states.

3.1. *Relationship between entropy rate and predictive information rate*

For a given size of state space N, the entropy rate can vary between zero for a deterministic sequence and $\log N$ for an uncorrelated sequence with $a_{ij} = 1/N$ for all i, j. Between these extremes, we find that the Markov chains that maximise the PIR have intermediate entropy. The scatter plot in fig. 3 was obtained by generating transition matrices at random by drawing each column independently from a Dirichlet distribution. (Matrices were drawn using several Dirichlet distributions with different parameters in order to cover the space more fully.) We also investigated optimising the PIR directly using a general purpose optimiser. We found that, for a range of different N, relatively sparse transition matrices maximise the PIR (see fig. 5). The 16×16 transition matrix is typical of what happens as N is increased: the conditional distribution for each antecedent state is approximately uniformly distributed across 3 or 4 states.

Fig. 6 shows a summary of these direct optimisations: for each N from 2 to 15, the optimiser was run on 15 different random initial conditions. The scatter plot shows the locally maximal PIR obtained from each of these, while the solid line shows the maximal entropy rate for each N, which is $\log N$. The results suggest that the maximal PIR for $N > 3$ is close to, but possibly slightly less than, $\frac{1}{2} \log N$; we have yet to prove this analytically. In most cases, the PIR is approximately equal to the entropy rate. In addition, not shown in the figure, the marginal entropies ($H(X)$ in our original notation) of all the optimised transition matrices are approximately $\log N$, indicating that the stationary distributions are close to uniform.

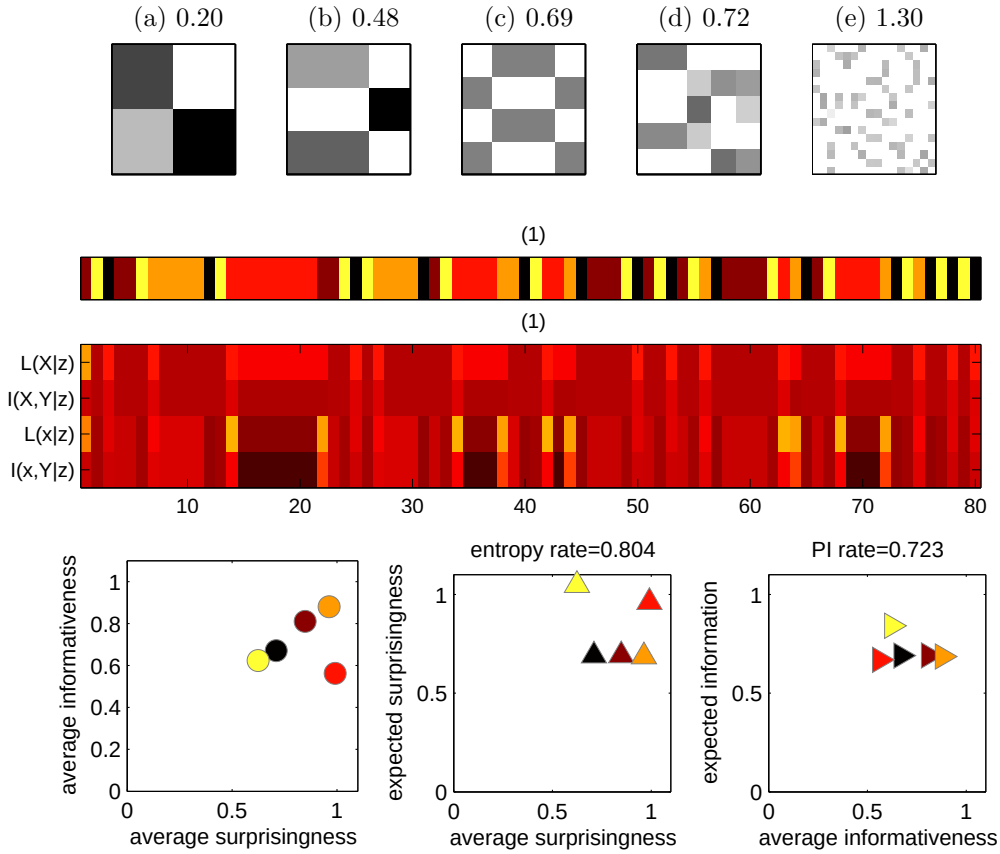

Figure 5. The results of direct numerical optimisation of the PIR for different state space sizes N. The number over each transition matrix is its PIR in nats/symbol ($1\,\text{nat} = \log_2 e \approx 1.44\,\text{bits}$). The panels below show a sample from transition matrix (d) and an information dynamic analysis as in fig. 4. The growth in maximal PIR with the number of states is a reasonable concommitant of the increase in maximal entropy rate $\log N$ with N.

4. Subjective information and model mismatch

In preceeding analysis, the surprisingness and predictive information were computed with respect to the observer's probabilistic model on the understanding that they represent the observer's subjective surprise and changes in beliefs on observing each symbol. The various averages used to obtain the entropy rate and predictive information rate were taken using the distributions implied by the model itself. When applied to some given sequence observations, these theoretical averages may or may not be close to the empirical average levels of surprise and information gain experienced by the observer as it processes the sequence. This will depend on whether or not the observer's model is a 'good' model of the data.

In some cases, such as when data is generated explicitly by sampling from some particular distribution, there can be said to be a 'true' model to which the observer's model can be compared. In others, the existence of a 'true' model may be questionable, or at least, not verifiable in practice. The most we can say in such cases is that one model may or may not be better than another according to certain criteria such as those advocated by Bayesians (24). In these cases, we take de Finetti at his word and say 'there are no real [i.e. objective] probabilities', only subjective ones (19).

However, returning to the Markov chain model we analysed in § 3, we can ask, what are the average levels of surprise etc. experienced when an observer using one Markov transition matrix processes a sequence generated from a Markov chain

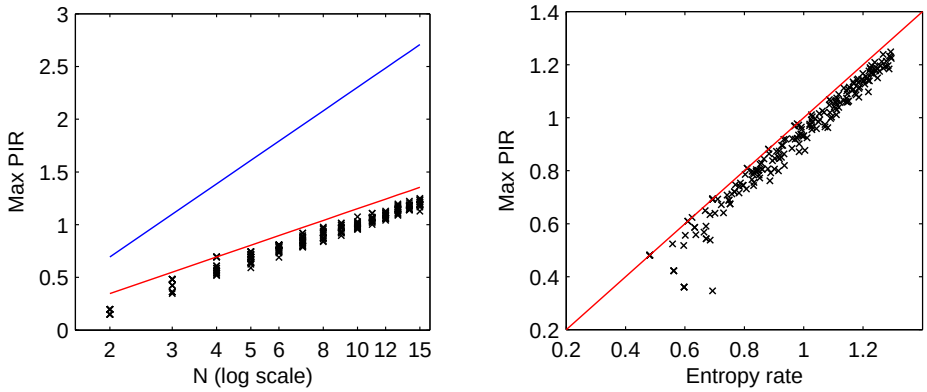

Figure 6. Results of optimising PIR for transition matrices with $2 \leq N \leq 15$. For each N, the optimiser was run with 15 different random intialisations. In the left-hand panel, the upper line is the maximal entropy rate for each N, i.e., $\log N$; the lower is at $\frac{1}{2}\log N$. The right-hand panel plots entropy rate against PIR for each of the optimised transition matrices; the diagonal line is where the two are equal.

using another transition matrix, assuming the state spaces are the same? That is, what if some or all of the averaging operations in equations (7) and (8) are carried out with respect to the true generative distribution rather than the observer's model? We will see that this leads to several variants of what was the entropy rate and predictive information rate, depending on how the averages are taken. In the following, we will assume an ergodic system so that we can equate these ensemble averages with the time averages that an observer could estimate given a long enough sample. Also, the assumption that the observer's state space includes at least all the symbols that can be generated requires a certain amount of prior knowledge on the part of the observer. More elaborate models based on, e.g., Dirichlet processes, can cope with an unknown and unbounded number of possible symbols, and will be useful to investigate in future.

4.1. *Surprise-based quantities*

We continue to denote the observer's probabilities with a p, but now introduce the generative probabilities with a q, e.g., $q_{X|Z}$, $q_{Y|X,Z}$ etc.. To parameterise the generative Markov chain, we introduce the generative transition matrix g.

The surprisingness of the observation $(X = x|Z = z)$ and the information it provides about the future are the same as in equations 1 and 2, since these are defined entirely in terms of the observer's subjective probability distributions. However, the average surprise following the observation $Z = z$ can now be computed using the actual distribution of symbols that occur after z as well as the distribution that the observer expects. Similarly, the average in-context suprisingness of symbol x as it actuall occurs can be computed using the true distribution $q_{Z|X=x}$. The general expressions for these averages are

$$\begin{aligned} \overline{\mathcal{L}}^*(z) &= \textstyle\sum_x \mathcal{L}(x|z) q(x|z), \\ \underline{\mathcal{L}}_*(z) &= \textstyle\sum_z \mathcal{L}(x|z) q(z|x), \end{aligned} \tag{9}$$

Note that $\overline{\mathcal{L}}^*(z) = \sum_x -q(x|z)\log p(x|z)$, the cross-entropy between the generative distribution $q_{X|Z=z}$ and the predictive distribution $p_{X|Z=z}$, and is therefore lower-bounded, for any z, by the entropy of the generative distribution,

$\sum_x -q(x|z) \log q(x|z)$. For the Markov chain in particular, we obtain

$$\begin{aligned} \overline{\mathcal{L}}^*(j) &= \textstyle\sum_{i=1}^N -g_{ij} \log a_{ij}, \\ \underline{\mathcal{L}}_*(i) &= \textstyle\sum_{j=1}^N -g^\dagger_{ij} \log a_{ij}. \end{aligned} \tag{10}$$

By averaging over both variables using either the observer's or the generative models, we can obtain several variants of the entropy rate in addition to the standard definition. Most of these do not have any obvious interpretation, but the following are suggestive of quantities that might be relevant to an observer:

$$\begin{aligned} \overline{\underline{\mathcal{L}}}_* &= \textstyle\sum_{x,z} \mathcal{L}(x|z) p(x|z) q(z), \\ \overline{\underline{\mathcal{L}}}^*_* &= \textstyle\sum_{x,z} \mathcal{L}(x|z) q(x,z). \end{aligned} \tag{11}$$

The first of these, $\overline{\underline{\mathcal{L}}}_*$ is the average level of uncertainty about the next symbol experienced by the observer while processing sequences from the generative model. The second, $\overline{\underline{\mathcal{L}}}^*_*$ is the average level of surprise experienced by that observer, and is bounded from below by the entropy rate of the generative model. The two are distinct: it is possible for an observer with bad model to be very certain, but wrong, about each coming symbol and thus be continually surprised. Conversly, the observer's model could make very broad, uncertain predictions and yet always find that the most likely predicted symbol appears. In the Markov chain, these measures evaluate to

$$\begin{aligned} \overline{\underline{\mathcal{L}}}_* &= \textstyle\sum_{i,j} \mathcal{L}(i|j) a_{ij} \pi^g_j, \\ \overline{\underline{\mathcal{L}}}^*_* &= \textstyle\sum_{i,j} \mathcal{L}(i|j) g_{ij} \pi^g_j, \end{aligned} \tag{12}$$

where $\mathcal{L}(i|j) = -\log a_{ij}$ and π^g is the stationary distribution resulting from the generative transition matrix g.

4.2. *Predictive information-based quantities*

We can repeat the same process of taking averages with respect to the generative distributions using the predictive information $\mathcal{I}(x|z)$ instead of the surprisingness $\mathcal{L}(x|z)$. For the remainder of this section, we give the results for the Markov chain in parallel with the general expressions, with $\mathcal{I}(i|j) = \sum_k a_{ki}(\log a_{ki} - \log[a^2]_{kj})$.

In addition to the quantities already defined in (8), we obtain two variants of the symbol-specific average predictive information measures.

$$\begin{aligned} \overline{\mathcal{I}}^*(z) &= \textstyle\sum_x \mathcal{I}(x|z) q(x|z), \qquad & \overline{\mathcal{I}}^*(j) &= \textstyle\sum_i \mathcal{I}(i|j) g_{ij}, \\ \underline{\mathcal{I}}_*(x) &= \textstyle\sum_z \mathcal{I}(x|z) q(z|x), \qquad & \underline{\mathcal{I}}_*(i) &= \textstyle\sum_j \mathcal{I}(i|j) g^\dagger_{ij}. \end{aligned} \tag{13}$$

These are the average predictive information gained after z and the average informativeness of x. Of the global measures of predictive information rate, there are again two new variants that are readily interpretable:

$$\begin{aligned} \overline{\underline{\mathcal{I}}}_* &= \textstyle\sum_{x,z} \mathcal{I}(x|z) p(x|z) q(z), \qquad & \overline{\underline{\mathcal{I}}}_* &= \textstyle\sum_{i,j} \mathcal{I}(i|j) a_{ij} \pi^g_j, \\ \overline{\underline{\mathcal{I}}}^*_* &= \textstyle\sum_{i,j} \mathcal{I}(i|j) g_{ij} \pi^g_j, & \overline{\underline{\mathcal{I}}}^*_* &= \textstyle\sum_{x,z} \mathcal{I}(x|z) q(x,z). \end{aligned} \tag{14}$$

If $\overline{\mathcal{I}}(z) = \sum_x \mathcal{I}(x|z)p(x|z)$ is thought of as the 'information expectancy' engendered by the context $Z = z$, then $\underline{\overline{\mathcal{I}}}_*$ is the average information expectancy experienced by observer while processing the data. In contrast, $\underline{\overline{\mathcal{I}}}^*_*$ is the average information actually received per symbol as measured by changes in beliefs about the future of the sequence.

4.3. *Effects of learning Markov chains*

Using the subjective information measures defined above, we can examine how an observer's assessment of a sequence drawn from a Markov chain changes as the observer gradually modifies its transition matrix to match that of the generative process. If the observer's model converges to the true transition matrix, then its estimates of the entropy and predictive information rates of the process will converge to those of the generative model, but depending on the observer's initial model and the learning process, these estimates will follow a certain trajectory. In particular, the observer's subjective predictive information rate may increase or decrease in response to adaptive learning. The converse does not follow automatically: convergence of the entropy and predictive information rates does not imply convergence of the transition matrix, but if the generative system is indeed ergodic, then all parts of the state space will be visited eventually with positive probability. In this case, the learning algorithm, which is, after all, just a matter of counting transitions, will almost certainly converge to the true transition matrix in the long run, as long as the 'forgetting rate' (see below) is set to zero.

We take a Bayesian approach to learning Markov chain parameters from observations: the observer's beliefs are represented by a distribution over possible transition matrices, which is updated using Bayes' rule after each symbol is observed. The algorithm is a direct generalisation of well known method for inferring a discrete distribution assuming a Dirichlet prior; we simply estimate one discrete distribution for each of the N possible antecedent states.

We can allow for the possibility that the transition matrix might change over time by broadening the current distribution over transition matrices between each observation, approximating a model in which the transition matrix follows a random walk. In practice, this means that the system is able to 'forget' about the distant past and estimate a transition matrix fitted to more recent observations.

For computational convenience we represent the observer's beliefs about the transition matrix with a product of Dirichlet distributions, one for each column of the transition matrix, that is,

$$p(a|\theta) = \prod_{j=1}^{N} p_{\mathrm{Dir}}(a_{:j}|\theta_{:j}), \tag{15}$$

where $a_{:j}$ is the j^{th} column of a and θ is an $N \times N$ matrix of parameters such that $\theta_{:j}$ is the parameter tuple for the N-component Dirichlet distribution p_{Dir},

$$p_{\mathrm{Dir}} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}, \qquad p_{\mathrm{Dir}}(\alpha|\varphi) = \frac{1}{B(\varphi)} \prod_{i=1}^{N} \alpha_i^{\varphi_i - 1}, \tag{16}$$

where $B : \mathbb{R}^N \to \mathbb{R}$ is the multinomial Beta function.

Ideally, the 'forgetting step' would be modelled by formulating a conditional distribution for the transition matrix given the transition matrix at the previous time point, that is, a random walk over transition matrices. This would in turn deter-

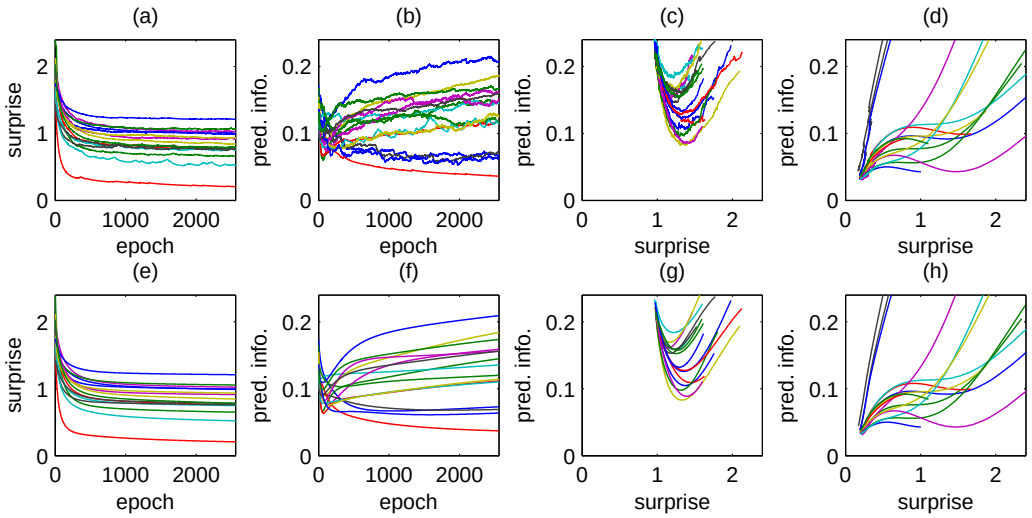

Figure 7. Learning dynamics in adaptive Markov chain system. The upper row shows the actual stochastic learning while the lower shows the idealised deterministic learning (see text accompanying eq. (19) for an explanation of how these were obtained).
Plots (a/b/e/f) show multiple runs starting from the same initial condition but using different generative transition matrices. Plots (c/d/g/h) show multiple runs starting from different initial conditions and converging on two transition matrices with (c/g) high or (d/h) low PIR respectively.
In (a/e) the average subjective surprisingness tends to decrease as it should since this the objective of learning. In (b/f), the subjective predictive information rate, after an initial transient phase, can go up or down depending on the generative system. The two target systems in (c/g) and (d/h) correspond to the highest and lowest lines in (a/b/e/f).

mine how a *distribution* over transition matrices should evolve under the action of the random walk, that is, a diffusion process. Unfortunately, direct attempts to construct such a model compatible with the use of Dirichlet distributions to represent the current belief state are not fruitful. Instead, we simulate a diffusion process at each time step by updating the Dirichlet parameters under the mapping

$$\theta_{ij} \mapsto \frac{\theta_{ij}}{1 + \beta\theta_{ij}}, \tag{17}$$

where β is a parameter which controls the forgetting rate. This tends to broaden or spread out the distribution, as would a true diffusion. Other mappings could also reasonably model the effect of a random walk and could be chosen here instead; for example, the mapping

$$\theta_{ij} \mapsto \zeta_{ij} + \frac{\theta_{ij} - \zeta_{ij}}{1 + \beta(\theta_{ij} - \zeta_{ij})}. \tag{18}$$

would cause the parameter matrix to evolve towards a 'background' state represented by ζ, and could simulate a random walk biased towards a certain point.

After the 'forgetting' step, the next observed symbol provides fresh evidence about the current transition matrix, which enables the observer to update its belief state. The choice of the Dirichlet distribution (being the conjugate prior of the multinomial distribution) makes these updates particularly simple: on observing the symbol i following symbol j, we increment θ_{ij} by 1. As a check, this stochastic online learning can be compared with what we would expect on average by replacing

the single element increment with the mapping

$$\theta \mapsto \theta + g\pi^g. \tag{19}$$

This simulates the simultaneous observation of all possible transitions weighted by their relative probabilities.

We applied the above system using many combinations of generative transition matrix g and initial observer state θ, both drawn at random. In each case, the evolution of the average surprisingness $\overline{\underline{\mathcal{L}}}^*_*$ and the predictive information rate $\overline{\underline{\mathcal{I}}}^*_*$ was recorded. Some of the aggregate results are shown in fig. 7, with stochastic learning in the upper row of plots and simulated deterministic learning in the lower row.

Notably, we find that, starting from a fixed belief state and depending on the statistics of the generative system, the subjective predictive information rate can increase or decrease as learning proceeds. This sort of behaviour is suggestive of the effect mentioned in § 2.4, where human subjects change their assessment of aesthetic value after repeated exposure to a stimulus, except that in this case, we are not *repeating* the same stimulus, but exposing the system to a prolonged sample from the generating process.

5. Experiments with minimalist music

Returning to our original goal of modelling the perception of temporal structure in music, we computed dynamic information measures for two pieces of minimalist music by Philip Glass, *Two Pages* (1969) and *Gradus* (1968). Both are monophonic and isochronous, and so can be represented as a sequence of symbols where each symbol stands for one note and time maps identically onto position in the sequence. Hence, the pieces can be represented very simply yet remain ecologically valid examples of 'real' music.

Music in the minimalist style was specifically chosen because, more than other styles of music, minimalist music tends to be constructed around patterns that are introduced in each piece as it develops, and rather less on pre-existing conventions and stylistic norms. In terms of the present discussion, we would say that in a minimalist piece, the significant expectations arise from the apprehension of regularites observed in that piece as it develops, relying less on statistical regularities representative of a particular style, such as Baroque music or Blues, which a listener must have previously internalised to fully appreciate such styles. To put it more succiccintly, the composition relies more on *intra*- as opposed to *extra*-opus stylistic norms. This means that, in our analysis, we can start with a 'vanilla' model, which, though capable of learning, does not initially embody any stylistic expectations such as might be gained by training on a corpus of music in a particular style.

Having said that, it should be noted that *Two Pages* embodies this ideal more than *Gradus*, the latter relying somewhat on expectations generated by familiarity with tonal music (36).

5.1. *Methods*

Since the aim of this experiment was not to find the best fitting model of the music, but rather to examine the behaviour of the dynamic information measures, we used the adaptive Markov chain model analysed in § 3 and § 4.3. Whilst Markov chains are not necessarily good models of music, using them does keep the computation

of the various information measures relatively simple.

For *Two Pages*, the distibution spreading map (17) was used with β set to 0.0004 and the Dirichlet parameters θ_{ij} initialised to 0.3 for all i, j. For *Gradus*, the parameters were roughly hand optimised to minimise the mean of the surprise under the constraint that all elements of θ be initialised to the same value. The mean surprisingness reached 1.29 nats/symbol with $\beta = 0.09$ and all entries of θ initialised to 0.02. By way of comparison, the multiple viewpoint variable order Markov model applied to *Gradus* in (36) acheived an average of 1.56 nats/symbol. Both of these are much lower than the $\log 12 = 2.48$ nats/symbol that would be obtained with a naïve encoding of the sequence (there are 12 symbols including one for rests), and less than the 2.29 nats/symbol that would be obtained using an encoding based on the marginal distribution of symbols.

The initial Dirichlet parameters have relatively little effect on the results, apart from during an initial transient phase, whereas the adaptation rate β affects the relative prominence of variations in the information measures in response to local variations as compared with those in response to larger changes in between sections: when β is small (slow adaptation), local features within a section are relatively less pronounced.

For the sake of comparison, we applied two rule-based analysis methods to both *Two Pages* and *Gradus*. Lerdahl and Jackendoff's grouping rules (37) are based Gestalt principles (38) applied to pitched events; we used an implementation of Grouping Preference Rule 3a (GPR3a). Cambouropoulos' Local Boundary Detection Model (39) is a later implementation of Gestalt principles, and for our purposes, can be considered to supercede Lerdahl and Jackendoff's rules. Both analyses result in a continuous-valued signal that can be interpreted as a 'boundary strength', that is, the degree to which a boundary (phrase, sectional, metrical) is indicated at each point in time.

In addition (but bearing in mind the caveats of § 2.5) we also computed several functions of the observer's time-varying model, which in this case consists of the expectation of the current Dirichlet distribution over transition matrices. If the Dirichlet parameters at a given time are $\theta \in \mathbb{R}^{N \times N}$, then the elements of the expected transition matrix a at that time are $a_{ij} = \theta_{ij} / \sum_k \theta_{kj}$. The functions computed were the entropy rate $\dot{\mathcal{H}}(a)$, the predictive information rate $\dot{\mathcal{H}}(a^2) - \dot{\mathcal{H}}(a)$, and the redundancy $\mathcal{H}(\pi^a) - \dot{\mathcal{H}}(a)$.

The redundancy was included as it corresponds with the 'information rate' (IR) that Dubnov *et al* (21, 22) compute for audio signals, in that it is the mutual information between past and present *for the currently estimated model*. In their experiments, Dubnov *et al* use a model which assumes that the frequency bands of the log-magnitude spectrogram of the audio signal are linear combinations of independent Gaussian processes. This model is fitted to blocks (called 'macro-frames' in the original papers) of audio data using SVD and standard power-spectral estimation methods. For reasons which we touched upon in § 2.6 and which we will discuss further in § 6, there are some ambiguities and inconsistencies in this application of the IR. However, using it, Dubnov *et al* find a correlation between the IR and continuous ratings of 'emotional force' made by human listeners.

5.2. *Results*

5.2.1. Two Pages

Traces of some of the dynamic information measures are shown in fig. 8 and fig. 11, along with some structural information about the pieces. In the case of *Two*

Two Pages, information dynamic analysis

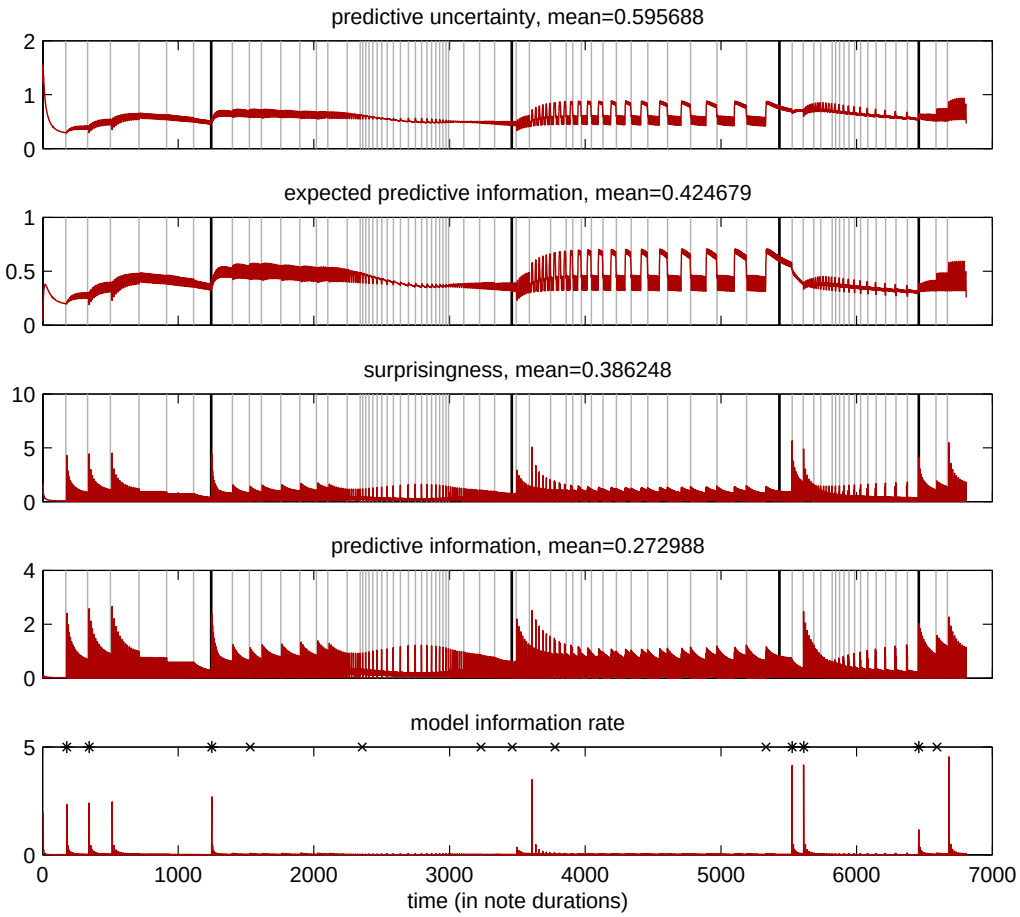

Figure 8. Analysis of *Two Pages*. In all panels, the thick vertical lines indicate the part boundaries as indicated in the score by the composer. The thin grey lines in the top four panels indicate changes in the melodic 'figures' of which the piece is constructed. In the bottom panel, the black asterisks indicate the six most surprising moments selected by Keith Potter, while the black crosses indicate an additional seven significant moments chosen by Potter at a later time. All information measures are in nats.

Two Pages, information dynamic analysis of model only

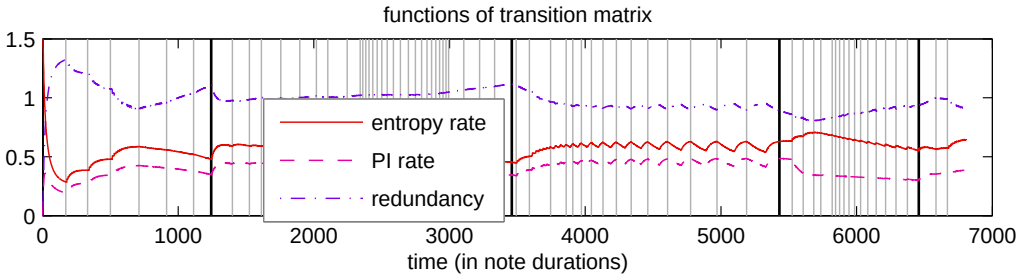

Figure 9. Information dynamic analysis of *Two Pages* using functions of the currently estimated transition matrix only. The redundancy as plotted here is the mutual information between the past and the present *given* the currently estimated model, that is $\mathcal{H}(\pi^a) - \dot{\mathcal{H}}(a)$, where $\mathcal{H}(\pi^a)$ is the entropy of the stationary distribution and a is the (time varying) estimated transition matrix.

Two Pages, rule based analysis

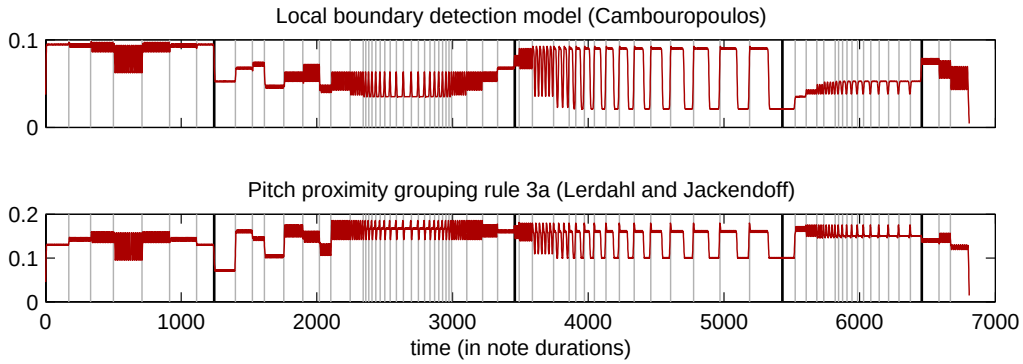

Figure 10. Analysis of *Two Pages* using (top) Cambouropoulos' Local Boundary Detection Model (LBDM) and (bottom) Lerdahl and Jackendoff's grouping preference rule 3a (GPR3a), which is a function of pitch proximity. Both analyses indicate 'boundary strength'.

Pages, the correspondence between the information measures and the structure of the piece is very close. In particular, there is good agreement between the model information signal (bottom panel) and the six 'most surprising moments' (marked with asterisks on the same plot) that music theorist and expert on minimalist music Keith Potter was asked to choose in a previous analysis of the piece (36). What appears to be an error in the detection of a major part boundary—between events 5000 and 6000 in fig. 8—actually raises a known anomaly in the score, where Glass places the boundary several events before there is any change in the pattern of notes. Alternative analyses of *Two Pages* place the boundary in agreement with the peak in our surprisingness signal.

At a later date, we asked Potter to select an unspecified number of additional significant moments. He chose eight points, seven of which are marked with crosses in fig. 8 (one was the end of the piece, which falls outside the scope of our analysis). While the initially selected six moments agree quite well with the model information signal, the subsequently selected seven are not well predicted by our analysis. If we refer to these as x_1–x_7, then x_7 matches exactly with a peak in the surprisingness and predictive information signals; x_1 is one note before a moderately sized peak, and the other five match local peaks which do not stand out from neighbouring peaks. An examination of the score, and Potter's own explanations for his choices, suggest that the Markov chain model is too weak to detect the significance of these points. For example, some of the changes involve transitions from regular repetition of a melodic figure to an expanding or contracting number of repetitions, the archetype being something like

$$\cdots a\ b\ b\ b\ a\ b\ b\ b\ a\ b\ b\ b\ a\ b\ b\ b\ \underline{b}\ a\ b\ b\ b\ b\ b\ a\ b\ b\ b\ b\ b\ b\ \cdots$$

To a human, having detected a periodicity of 4 in the initial segment, the first instance of a fourth consecutive b (underlined) is surprising and marks the start of an expansion process, but to a Markov model, the transition (b a) is always more surprising than the transition (b b): loosely speaking, first order Markov chains cannot 'count'. Clearly, the human listener is bringing more sophisticated machinery to bear than the humble Markov chain.

Elsewhere in the analysis, there is some noticable structure in the predictive uncertainty and expected predictive information signals in the third section, where there is a clear alternation between two levels in each of the gradually lengthening figures. The first part of each figure consists of a pattern similar to that which

opens the piece, while the second half consists of a repeating pattern of five notes using three pitches. Over multiple alternations between the two patterns, the model adopts a transition matrix which attaches different levels of uncertainty to each pitch. The pattern in the second half of each of these figures uses the uncertainty-creating notes more often and thus creates a higher level of average uncertainty.

Fig. 9 shows an analysis derived by considering the currently estimated transition matrix only. Since this varies relatively slowly according to the forgetting rate β, this type of analysis does not tend to produce sharply localised responses to individual events. In particular, the major boundaries in the piece are not as clearly indicated as in the bottom three panels of fig. 8. The entropy rate and PIR are approximately smoothed versions of the predictive uncertainty and the expected predictive information respectively.

The results of applying the two rule-based methods are shown in fig. 10. We note that while both methods obviously reflect the structure of the piece, there is no sense of a hierarchy of large and small peaks to indicate major and minor boundaries. Indeed, many of the boundaries which are very clear in the lower three panels of fig. 8, including all of the main sectional boundaries, do not produce peaks at all in the rule-based analyses—the perception of these boundaries seems to depend entirely on the interaction between the events and the dynamically varying model, rather than any general (but static) principles of tonal music, which the rule-based analyses might be said to embody.

5.2.2. Gradus

Gradus is much less systematically structured than *Two Pages*, and relies more on the conventions of tonal music, which are not represented the model. The information dynamic analysis shown in fig. 11 does not have such a transparent interpretation as that of *Two Pages*; nonetheless, there are many points of correspondence between the analysis and the segmentation given by Keith Potter (36).

For example, peaks in the model information rate at bars 42 and 66 mark major sectional boundaries in the piece. The boundary at bar 83 is less marked in the model information signal but clearly visible in the surprisingness and predictive information signals. The major sections are visible in the broad arch-shaped developments of the predictive uncertainty signal.

Peaks at bars 4, 21, 23, 50, 60, 66, 71 and 87 all coincide with the introductions of new pitches, or in some cases, the re-introduction of a pitch that had been absent for some time. The peak at bar 44 marks the noticable introduction of a rising sharp fourth from C♯ to G. Other peaks, such as those at bars 6, 7, 17, 19, 26 and 68, are related to changes in the melodic pattern, or in the case of the peak in the surprisingness signal around bar 35, a switch to a more fragmented rhythm with greater numbers of rests.

Towards the end of the piece, the peak at 94 marks the very prominent first occurence of the repeated notes which bring the piece to a close, after which all the information measures begin to tail-off. See (36) for further musicological analysis of *Gradus*.

In addition to these traces of the overall structural development, there appear to be correlations between the predictive uncertainty signal and the perception of rhythmic and metrical stress. In several places where there is a strong sensation of duple time (e.g., the end of bar 7, the first half of bar 18, and in bars 72, 75 and 80), there is an accompanying pattern of alternating high and low predictive uncertainty. This is not visible in the figure because it occurs at the level of individual notes, but we observe that the stress is felt on the note *following* the note which creates higher uncertainty.

Furthermore, evidence of the 32-quaver metre can be found by computing the

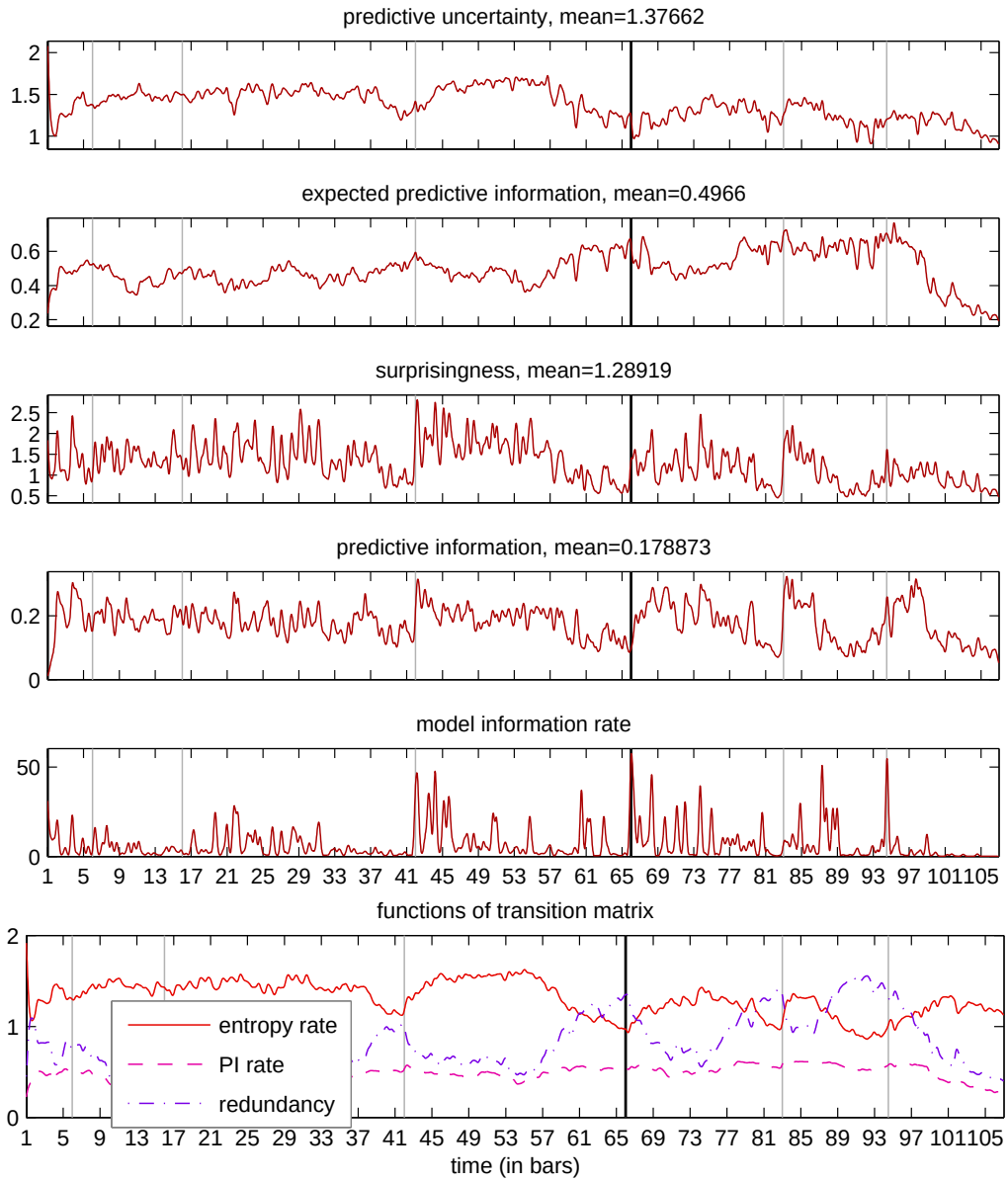

Figure 11. Analysis of *Gradus*. In all panels, the thick black vertical lines indicate the part boundaries as indicated in the score by the composer. The thin grey lines indicate a segmentation given by Keith Potter. Note that the traces were smoothed with a Gaussian window about 12 events wide to make them more legible. All information measures are in nats.

averages of the dynamic information signals for notes at each of the 32 metrical positions. The first note of each bar is, on average (with respect to this particular model), more surprising (by approximately 0.9 nats), and more informative (by about 0.1 nats) than the other metrical positions, all of which are roughly equal according to both measures (see fig. 13). There is some evidence of a 64-quaver hypermetre, but when the analysis is performed at a 128-quaver level, the dominant periodicity appears to remain at 64 quavers. Not shown in the figure, the predictive uncertainty tends to be highest before the first note of each bar and lowest after the first note of each bar.

The rule-based analyses of *Gradus* are illustrated in fig. 12. Boundaries at bars 42 and 83 are indicated in the LBDM analysis, but the major boundary at bar 66

Gradus, rule based analysis

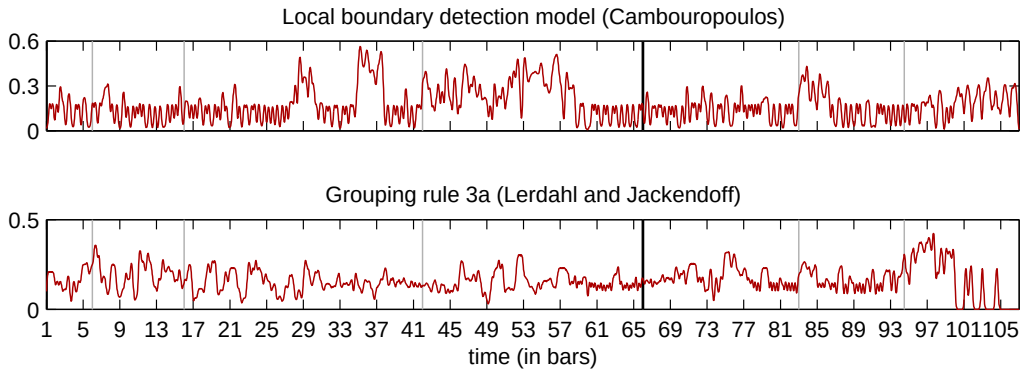

Figure 12. Boundary strength analysis of *Gradus* using (top) Cambouropoulos' (39) Local Boundary Detection Model and (bottom) Lerdahl and Jackendoff's (37) grouping preference rule 3a.

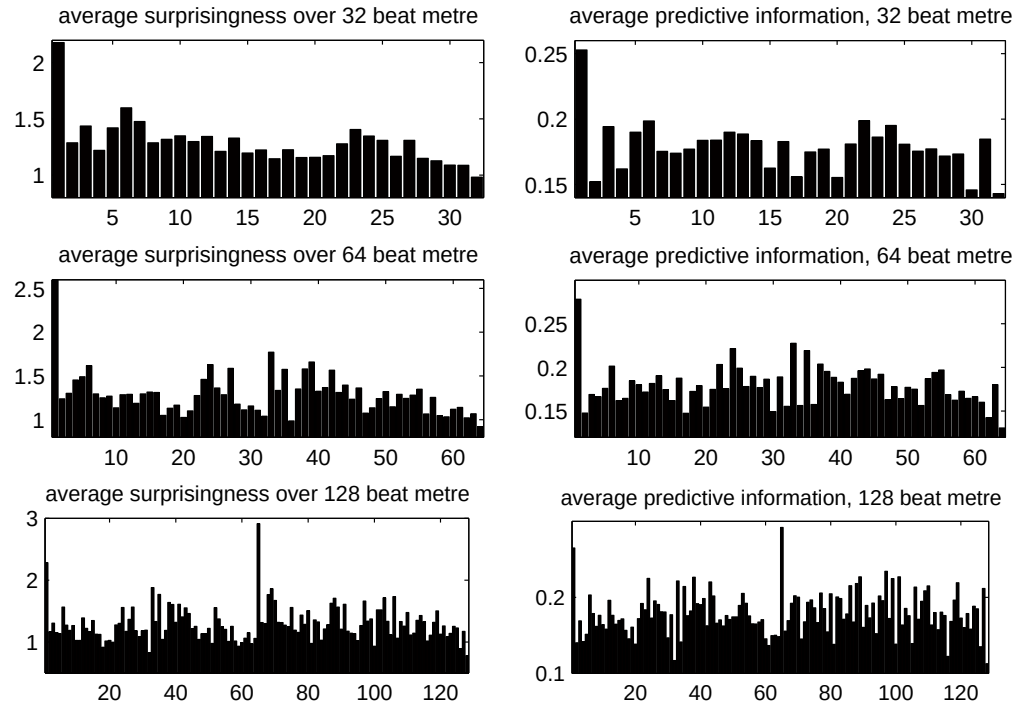

Figure 13. Information signals averaged over events at equivalent metrical positions assuming bar lengths of 32, 64, and 128 quavers. (The notated bar length is 32 quavers.)

is almost entirely absent. The prominent peaks at bars 29 and 35 coincide with the introduction of greater numbers of rests, but with respect to the overall structure of the piece, these peaks seem to be overstated. The GPR3a analysis responds to boundaries at bars 6 and 83, but not the major boundaries at bars 42 and 66.

6. Discussion and related work

Our definitions of the predictive information and predictive information rate are distinct from the predictive information of Bialek *et al* (29). They too consider stationary random processes, but proceed by examining the entropy of a segment of finite duration T, which, given the assumption of stationarity, will be a function of T alone, say $\mathcal{S}(T)$. This entropy will increase with increasing T, tending towards

a linear growth at a rate equal to the entropy rate of the process. The mutual information between two adjacent segments, of duration T and T' respectively, can be expressed in terms of $\mathcal{S}$. Bialek *et al* define the predictive information as the limit of this as T' tends to infinity:

$$I_{\mathrm{pred}}(T) = \lim_{T' \to \infty} \mathcal{S}(T) + \mathcal{S}(T') - \mathcal{S}(T + T'). \tag{20}$$

As T increases, $I_{\mathrm{pred}}(T)$ may tend to a finite limit (which might be zero) or increase indefinitely, tending to logarithmic or fractional power-law growth. The type of growth characterises a fundamental aspect of the stochastic complexity of process, and will be very interesting to study further, especially in relation to characterising long-term dependencies in music. Though I_{pred} certainly fits naturally into the information dynamics 'toolbox', we would argue that such as the ones we describe should also be considered, since $I_{\mathrm{pred}}(T)$ is a *global* measure which applies to the random process as a whole, not to specific realisations, much less to specific instants within a realisation.

The idea of measuring information gained about model parameters as the KL divergence between prior and posterior distributions is equivalent to Itti and Baldi's 'Bayesian surprise' (27). As we noted in § 2.5, we suspect that the real significance of this information about *parameters* is its indirect effect on the observer's expectations about *future observables*, i.e., its status as a form of predictive information.

Dubnov's 'information rate' (IR) (21) is a measure of redundancy defined as the mutual information between the past and the present, or in the notation of § 2, $I(Z, X)$. We discussed in § 2.6 some of the issues surrounding the interpretation of this definition. Here, we will examine the claim that the IR exhibits the kind of 'inverted-U' behaviour we obtain with our predictive information rate when applied to Markov chains § 3.1.

In general, the IR of a random process is $I(Z, X) = H(X) - H(X|Z)$. For a noise-like process with no temporal dependencies, $H(X|Z) = H(X)$ and the IR is indeed zero. Dubnov states that deterministic processes also have low IR, by arguing that in such processes $H(X)$, which is an upper bound on the IR, must be low. However, it is quite easy to design processes that have high marginal entropy $H(X)$, but become completely determined on observing only a few elements of the sequence, e.g., a constant signal with an initially unknown value distributed uniformly over a wide range. This shows that $H(X)$ can be large even when $H(X|Z)$ is zero; it is these processes which maximise the IR, not those of intermediate randomness.

Similarly, the IR of a Markov chain with transition matrix a is $\mathcal{H}(\pi^a) - \dot{\mathcal{H}}(a)$. As we saw in § 3.1, the PIR is maximised by Markov chains of intermediate entropy rate. In contrast, the IR is maximised by a Markov chain having a uniform initial/equilibrium state distribution π^a but which cycles deterministically through all states thereafter. One would expect such a predictable sequence to be rather uninteresting: the PIR is zero but, with N available states, the IR takes the maximum possible value of $\log N$. Dubnov's expression for the IR in Gaussian processes exhibits similar behaviour: in this case the spectral flatness measure (SFM) is shown to vary inversely with the IR, but the processes which minimise SFM (and therefore maximise IR) are those with maximally sparse power spectra, i.e. sinusoidal waveforms which are indefinitely predictable once the phase and amplitude have been fixed by observing two samples.

Eerola *et al* (8) propose a similar approach to ours, emphasising the need for dynamic probability models when judging uncertainty and predictability of musical patterns. They also describe experimental methods for assessing these quantities in human listeners. However, they do not explore the possibilites for multiple in-

formation measures or consider the concept of predictive information.

In § 5, we applied the Markov chain model to pieces of music chosen partly because they are monophonic and isochronous, enabling a straightforward encoding of each note or rest as a state in a Markov chain. Though quite restrictive, these constraints are compatible with those defined by Larson for the '*Seek Well* creative microdomain' (40, 41). Larson asked human subjects to invent continuations of short melodic fragments under the same constraints and found remarkable agreement among subjects given the combinatorial explosion of possibilities as the length of the generated sequence increases. We may be able to apply our Markov chain-based information dynamic analysis to this data.

Our approach to the perception of musical structure is very much in the same spirit as that of David Huron as expounded in his book *Sweet Expections* (42). Huron reports and summarises many experiments showing that human subjects do indeed behave as if they have internalised the statistical regularities present in music, and also discusses how and why the perception of statistical structure (Berlyne's collative variables again) might be closely related to affect and emotional response. For example, Huron suggests that the *qualia* induced by different scale degrees in a tonal setting might be explained in terms of statistical or collative properties which are essentially equivalent to the per-state information dynamic quantities defined in § 2, such as $\overline{\mathcal{L}}(z)$, the average uncertainty engendered by a pitch, and $\underline{\mathcal{L}}(x)$, whether or not a pitch tends to be surprising when it occurs.

7. Conclusions and future work

We have described an approach to the analysis of temporal structure based on an information-theoretic assessment made from the point of view of an observer that updates its probabilistic model of a process dynamically as events unfold. In principle, any dynamic probabilistic model can be given this treatment. In this paper, we have examined the information dynamics of Markov chains, found an intruiging inverted-'U' relationship between the entropy rate and the PIR, and applied the method to the analysis of minimalist music, with some encouraging results but raising many questions and suggesting several possible developments.

Firstly, we would like to extend the analysis to more complex models such as those involving time-dependent latent variables, like HMMs, continuous valued-variables (e.g. Gaussian processes), and probabilistic grammars, such Bod's Data Oriented Parsing (43). The latter in particular, incorporating explicit tree structures over time, are better suited to modelling long-term dependencies and are likely to be a closer fit to the way humans process music.

Secondly, since pieces of music are relatively short compared with the amount of experience required to become familiar with musical styles, it will be necessary to collect models pre-trained on various style-specific corpora to act as the starting point for the processing of a particular piece. The combination of long and short term models has been investigated by Pearce and Potter *et al* (11, 36) and, with an application to motion capture data from dancers, by Brand (44).

Thirdly, to assess the cognitive relevance of our approach, we are planning experiments with human subjects to (a) search for physical correlates of the dynamic information measures, e.g. in EEG data, and (b) determine whether or not there is any relationship between the predictive information rates and the subjective experience of 'interestingness' and aesthetic value.

Fourthly, though we have not touched on variation in event durations, this is likely to be an important aspect of the information dynamics of music, since rhythm is fundamental to music. The simple fact that usually we do not know how a long

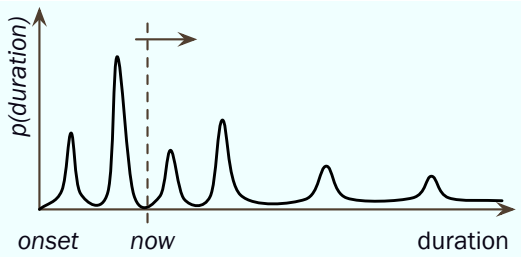

Figure 14. Illustration of how information about an unknown duration accumulates as the duration unfolds in real time. Depending on the observer's beliefs about the distribution of durations, information can arrive at a non-uniform rate while the observer is waiting for the duration to expire.

a note will be while it is sounding implies that we are receiving information (about the duration) even while, ostensibly, nothing is happening. The information rate will depend on the observer's probability distribution over possible note durations, as illustrated in fig. 14.

Fifthly, while generating sequences from different Markov chains in the course of this work, we became aware that it can be mildly entertaining to select Markov chains for sonification by generating large numbers of transition matrices (e.g. by sampling from a Dirichlet distribution) and automatically selecting the few with the highest predictive information rates. Two or three such sequences played in parallel, perhaps in different pitch ranges and at different rates, can, we dare say, be quite musical. These experiences suggest possible applications in computer-assisted composition—the sequences generated are not pieces of music by any stretch, but they can provide musical material. Indeed, the process parallels what seems to occur when humans compose, in that generative but relatively uncritical phases are interleaved with selective phases where aesthetic judgements are brought to bear (25). Related work includes Todd and Werner's surprise-driven model of bird song evolution (45) and Murray Brown's work on automatic composition (46).

In closing, we would like to cite some suggestive remarks from philosophers of music which have some resonance with what we are proposing. Davies (47) reviews a range of literature on musical affect under the heading of 'contour theories', which is meant to convey the notion of a curve in an abstract space with time along one axis and whose shape captures some structural essence of the music. For example, Langer (48) discusses a 'morphology of feelings', which operates at the level of 'patterns ... of agreement and disgreement, preparation, fulfilment, excitation, sudden change, etc.', arguing that these structures are relevant because they 'exist in our minds as "amodal" forms, common to both music and feelings.' Stern (49) used the term 'vitality effects' to describe 'qualities of shape or contour, intensity, motion, and rhythm—"amodal" properties that exist in our minds as dynamic and abstract, not bound to any particular feeling or event.' For example, 'bursting' could describe bursting into tears or laughter, a bursting watermelon, a burst of speed, a sforzando, and so on. Others examples include 'surging', 'fading', being 'drawn out' etc. Whilst such speculations are somewhat outside the scope of this paper, we do notice a common thread in the idea of an 'amodal' dynamic representation capturing patterns of change at an abstract level, something for which the information-dynamic approach may well provide a quantitative basis.

8. Acknowledgements

This research was supported by EPSRC grant GR/S82213/01. Thanks are also due to to Keith Potter and Geraint Wiggins (Goldsmiths, University of London) for providing musicological analyeses of *Two Pages* and *Gradus*, and to Marcus Pearce (Goldsmiths, University of London) for providing the rule-based analyses of both pieces.

References

[1] D. E. Berlyne. *Aesthetics and Psychobiology*. Appleton Century Crofts, New York, 1971.
[2] Leonard B. Meyer. *Music, the arts and ideas: Patterns and Predictions in Twentieth-century culture*. University of Chicago Press, 1967.
[3] Eugene Narmour. *Beyond Schenkerism*. University of Chicago Press, 1977.
[4] Eduard Hanslick. *On the musically beautiful: A contribution towards the revision of the aesthetics of music*. Hackett, Indianapolis, IN, 1854/1986.
[5] Richard T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13, 1946.
[6] Edwin T. Jaynes. How does the brain do plausible reasoning? In G. J. Erickson and C. R. Smith, editors, *Maximum-Entropy and Bayesian Methods in Science and Engineering*. Kluwer Academic, 1988.
[7] J. R. Saffran, E. K. Johnson, R. N. Aslin, and E. L. Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52, 1999.
[8] T. Eerola, P. Toiviainen, and C. L. Krumhansl. Real-time prediction of melodies: Continous predictability judgments and dynamic models. In C. Stevens, D. Burnham, G. McPherson, E. Schubert, and J. Renwick, editors, *Proceedings of the 7th International Conference on Music Perception and Cognition (ICMPC7)*, Syndney, Australia, 2002. Causal Productions.
[9] Daryl Conklin and Ian H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.
[10] D. Ponsford, G. A. Wiggins, and C. S. Mellish. Statistical learning of harmonic movement. *Journal of New Music Research*, 28(2):150–177, 1999. Also available as Research Paper 874, from the Division of Informatics, University of Edinburgh.
[11] Marcus T. Pearce. *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Department of Computing, City University, London, 2005.
[12] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656, 1948.
[13] Abraham Moles. *Information Theory and Esthetic Perception*. University of Illinois Press, 1966.
[14] J. E. Cohen. Information theory and music. *Behavioral Science*, 7(2):137–163, 1962.
[15] J. E. Youngblood. Style as information. *Journal of Music Theory*, 2:24–35, 1958.
[16] E. Coons and D. Kraehenbuehl. Information as a measure of structure in music. *Journal of Music Theory*, 2(2):127–161, 1958.
[17] Lejaren Hiller and Calvert Bean. Information theory analyses of four sonata expositions. *Journal of Music Theory*, 10(1):96–137, 1966.
[18] Daniel. E. Berlyne, editor. *Studies in the New Experimental Aesthetics: Steps towards an objective psychology of aesthetic appreciation*. Hemisphere, Washington D.C., 1974.
[19] Bruno de Finetti. *Theory of Probability*. John Wiley and Sons, New York, 1975.
[20] Scott J. Simon. *A Multi-dimensional Entropy Model of Jazz Improvisations for Music Information Retrieval*. PhD thesis, Univeristy of North Texas, 2005.
[21] Shlomo Dubnov. Spectral anticipations. *Computer Music Journal*, 30(2):63–83, 2006.
[22] Shlomo Dubnov, Stephen McAdams, and Roger Reynolds. Structural and affective aspects of music from statistical audio signal analysis. *Journal of the American Society for Information Science and Technology*, 57(11):1526–1536, 2006.
[23] Shlomo Dubnov. Unified view of prediction and repetition structure in audio signals with application to interest point detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):327–337, 2008.
[24] Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430), 1995.
[25] Margaret Boden. *The Creative Mind: Myths and Mechanisms*. Weidenfeld & Nicolson, 1990.
[26] S. Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society / Oxford University Press, 2001.
[27] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. In *Advances Neural in Information Processing Systems (NIPS 2005)*, 2005.
[28] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, Chichester, 1994.
[29] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2001.
[30] W. Wundt. *Outlines of Psychology*. Englemann, Lepzig, 1897.
[31] Leonard B. Meyer. Music and emotion: Distinctions and uncertainties. In Juslin and Sloboda (50), chapter 15, pages 341–360.
[32] Tod S. Levitt, Thomas O. Binford, and Gil J. Ettinger. Utility-based control for computer vision. In

Proc. Fourth Annual Conference on Uncertainty in Artificial Intelligence (UAI '88), pages 407–422, Amsterdam, 1990. North-Holland Publishing Co.

[33] James J. Gibson. *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston, 1966.

[34] Fred Attneave. Some informational aspects of visual perception. *Pyschological Review*, 61(3):183–193, 1954.

[35] Samer Abdallah. A critique of Dubnov's 'information rate'. Technical Report C4DM-TR-08-11, Centre for Digital Music, Queen Mary University of London, 2008. Available from `http://www.elec.qmul.ac.uk/~samer`.

[36] Keith Potter, Geraint A. Wiggins, and Marcus T. Pearce. Towards greater objectivity in music theory: Information-dynamic analysis of minimalist music. *Musicae Scientiae*, 11(2):295–324, 2007.

[37] Fred Lerdahl and Ray Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA, 1983.

[38] Wolfgang Köhler. *Gestalt Psychology*. Liveright, New York, 1947.

[39] Emilios Cambouropoulos. *Towards a General Computational Theory of Musical Structure*. PhD thesis, University of Edinburgh, 1998.

[40] Eric Nichols. Dynamic melodic expectation in the creative microdomain *Seek Well*. Technical Report CRCC Technical Report 138, Indiana University, USA, 2005.

[41] Steve Larson. Continuations as completions: Studying melodic expectation in the creative microdomain seek well. In Marc Leman, editor, *Music, Gestalt, and Computing: Studies in Cognitive and Systematic Musicology*, pages 321–334. Springer, Berlin, 1997.

[42] David Huron. *Sweet Expectations*. MIT Press, 2006.

[43] Rens Bod. A unified model of structural organization in language and music. *Journal of Artificial Intelligence Research*, 17:289–308, 2002.

[44] Matthew Brand and Aaron Hertzmann. Style machines. In Kurt Akeley, editor, *Siggraph 2000, Computer Graphics Proceedings*, pages 183–192. ACM Press / ACM SIGGRAPH / Addison Wesley Longman, 2000.

[45] Gregory M. Werner and Peter M. Todd. Too many love songs: Sexual selection and the evolution of communication. In *Fourth European Conference on Artificial Life*, pages 434–443, Brighton, England, 1997.

[46] Tim Murray Brown. An experiment with automated composition via expectation violation. Master's thesis, Computing Laboratory, University of Oxford, 2008. Available at `http://users.ox.ac.uk/~magd2227/project/`.

[47] Stephen Davies. Philosophical perspectives on music's expressiveness. In Juslin and Sloboda (50), chapter 2, pages 23–44.

[48] Susanne K. Langer. *Philosophy in a new key*. Harvard University Press, Cambridge, MA, 1957.

[49] D. Stern. *The Interpersonal World of the Infant*. Academic Press, London, 1985.

[50] Patrick N. Juslin and John A. Sloboda, editors. *Music and Emotion — Theory and Research*. Oxford University Press, 2004.

Appendix A. Derivations for Markov Chains

Let $S : \Omega \to \mathcal{A}^\infty$ be a random process whose realisations are infinite sequences of elements taken from an alphabet $\mathcal{A}$. The t^{th} element of the sequence is represented by the random variable $S_t : \Omega \to \mathcal{A}$. A realisation of the random process is a sequence $s = S(\omega) \in \mathcal{A}^\infty$, where ω is a sample drawn from a probability space on Ω. We assume that $\mathcal{A}$ contains N elements $\{\sigma_1, \ldots, \sigma_N\}$. If S is a Markov chain, then the process can be parameterised in terms of a transition matrix $a \in \mathbb{R}^{N\times N}$ and an initial distribution $b \in \mathbb{R}^N$ for the first element of the sequence:

$$a_{ij} \triangleq \Pr(S_{t+1} = \sigma_i | S_t = \sigma_j), \tag{A1}$$

$$b_i \triangleq \Pr(S_1 = \sigma_i). \tag{A2}$$

Note that $\Pr(\psi)$ denotes the probability that ψ is true where ψ is logical formula. Similarly, $\Pr(\psi|\phi)$ denotes the probability of ψ conditioned on the truth of ϕ. Probability distribution functions will be written as a p with a subscript to indicate from which random variables the arguments are intended to be drawn, e.g., $p_{S_t} : \mathcal{A} \to \mathbb{R}$ is the marginal distribution function of the t^{th} element of the chain and thus $p_{S_t}(\sigma_i)$ is the probability that S_t takes the value σ_i.

The equilibrium or stationary distribution $\pi^a \in \mathbb{R}^N$ of the Markov chain is defined by the condition $a\pi^a = \pi^a$, which implies that π^a is an eigenvector of the transition matrix a with eigenvalue 1. In order that the equilibrium distribution be unique, we require that the Markov chain be *irreducible*, i.e., that every state is potentially reachable from every other state, and also *aperiodic*. Together, these imply that the Markov chain is also *ergodic*.

Since we want the Markov chain to be stationary and to have a well defined equilibrium distribution π^a, we must have $\Pr(S_t = \sigma_i) = \pi^a_i$ for all t including $t = 1$. Hence, $b = \pi^a$, which is in turn a function of a.

A.1. *Entropy and entropy rate*

Having found the equilibrium distribution, we can derive the entropy of any single element S_t of the chain taken in isolation. For any t, $H(S_t) = \mathcal{H}(\pi^a)$, where $\mathcal{H}$ is the Shannon entropy function defined as

$$\mathcal{H} : \mathbb{R}^N \to \mathbb{R}, \qquad \mathcal{H}(\theta) = \sum_{i=1}^{N} -\theta_i \log \theta_i, \tag{A3}$$

The conditional entropy $H(S_{t+1}|S_t)$ can be derived by considering the joint distribution of S_{t+1} and S_t:

$$H(S_{t+1}|S_t) = \sum_{i=1}^{N}\sum_{j=1}^{N} -\Pr(S_{t+1}{=}\sigma_i \wedge S_t{=}\sigma_j) \log \Pr(S_{t+1}{=}\sigma_i | S_t{=}\sigma_j) \tag{A4}$$

Hence, we can define a function $\dot{\mathcal{H}} : \mathbb{R}^{N\times N} \to \mathbb{R}$ such that $H(S_{t+1}|S_t) = \dot{\mathcal{H}}(a)$:

$$\dot{\mathcal{H}}(a) \triangleq \sum_{i=1}^{N}\sum_{j=1}^{N} -a_{ij}\pi^a_j \log a_{ij} \tag{A5}$$

This is independent of t and yields the entropy rate of the process.

A.2. *Predictive information rates*

The predictive information rate (PIR), using X, Y, and Z to stand for the present, future and past respectively, can be written in several ways including

$$\begin{aligned} I(X,Y|Z) &= H(Y|Z) - H(Y|X,Z) \\ &= H(X|Z) - H(X|Y,Z). \end{aligned} \tag{A6}$$

At this point we will assume without loss of generality that the Markov chain extends infinitely in both directions and that the current time is zero, so that $Z = S_{-\infty:-1}$, $X = S_0$, and $Y = S_{1:\infty}$.

Now, in general, if three variables A, B and C are such that A and C are conditionally independent given B, that is, in the commonly understood abuse of notation, $p(c|b,a) = p(c|b)$, then $H(C|B,A) = H(C|B)$:

$$\begin{aligned} H(C|B,A) &= \sum_{a,b,c} p(c|b,a)p(b,a)\log p(c|b,a) \\ &= \sum_{b,c} p(c|b)\left(\sum_a p(b,a)\right)\log p(c|b) \\ &= \sum_{b,c} p(c|b)p(b)\log p(c|b) \\ &= H(C|B). \end{aligned} \tag{A7}$$

For the Markov chain, this implies that the PIR can be written as

$$\begin{aligned} I(S_0, S_{1:\infty}|S_{-\infty:-1}) &= H(S_{1:\infty}|S_{-\infty:-1}) - H(S_{1:\infty}|S_0, S_{-\infty:-1}) \\ &= H(S_{2:\infty}|S_1) + H(S_1|S_{-1}) - \big(H(S_{2:\infty}|S_1) + H(S_1|S_0)\big) \\ &= H(S_1|S_{-1}) - H(S_1|S_0) \end{aligned} \tag{A8}$$

The second term is the entropy rate $\dot{\mathcal{H}}(a)$ of the Markov chain, while the first term can be identified as the entropy rate of the Markov chain obtained by taking every second element of the original chain. The transition matrix of this derived two-step chain is simply the matrix square of the original transition matrix, i.e. a^2. If the original Markov chain is ergodic, the two-step chain will also be ergodic with the same equilibrium distribution, and the average predictive information rate will be $\dot{\mathcal{H}}(a^2) - \dot{\mathcal{H}}(a)$.

The predictive information for the Markov chain is derived by considering the information in the observation $S_0 = s_0$ about the entire tail of the sequence $S_{1:\infty}$ given the preceeding context $S_{-\infty:-1} = s_{-\infty:-1}$. We will write this as $I(S_0{=}s_0, S_{1:\infty}|S_{-\infty:-1}{=}s_{-\infty:-1})$ and compute it as the KL divergence between the prior $p_{S_{1:\infty}|S_{-\infty:-1}=s_{-\infty:-1}}$ and the posterior $p_{S_{1:\infty}|S_0=s_0,S_{-\infty:-1}=s_{-\infty:-1}}$. Because of the Markov dependency structure this can immediately simplified to

$$I(S_0{=}s_0, S_{1:\infty}|S_{-\infty:-1}{=}s_{-\infty:-1}) = D(p_{S_{1:\infty}|S_0=s_0}||p_{S_{1:\infty}|S_{-1}=s_{-1}}). \tag{A9}$$

Expanding this using the definition of the KL divergence (and dropping the subscripts of the distribution functions where the relevant random variables are clear

from the arguments) yields

$$
\begin{aligned}
&D(p_{S_{1:\infty}|S_0=s_0}||p_{S_{1:\infty}|S_{-1}=s_{-1}}) \\
&= \sum_{s_{1:\infty}\in\mathcal{A}^\infty} p(s_{1:\infty}|s_0)\log\frac{p(s_{1:\infty}|s_0)}{p(s_{1:\infty}|s_{-1}} \\
&= \sum_{s_{1:\infty}\in\mathcal{A}^\infty} p(s_{2:\infty}|s_1)p(s_1|s_0)\log\frac{p(s_{2:\infty}|s_1)p(s_1|s_0)}{\sum_{s_0'\in\mathcal{A}} p(s_{2:\infty}|s_1)p(s_1|s_0')p(s_0'|s_{-1})} \\
&= \sum_{s_1\in\mathcal{A}}\left(\sum_{s_{2:\infty}\in\mathcal{A}^\infty} p(s_{2:\infty}|s_1)\right)p(s_1|s_0)\log\frac{p(s_1|s_0)}{\sum_{s_0'\in\mathcal{A}} p(s_1|s_0')p(s_0'|s_{-1})} \\
&= \sum_{s_1\in\mathcal{A}} p(s_1|s_0)\log\frac{p(s_1|s_0)}{\sum_{s_0'\in\mathcal{A}} p(s_1|s_0')p(s_0'|s_{-1})}
\end{aligned}
\tag{A10}
$$

This shows that the information in $S_0{=}s_0$ about the entire future is accounted for by information it contains about the next element of the chain. Rewritten in terms of the transition matrix, the predictive information is a function of the current and previous states alone:

$$
\begin{aligned}
\mathcal{I}(i|j) &= I(S_0{=}\sigma_i, S_1|S_{-1}{=}\sigma_j) \\
&= \sum_{k=1}^{N} a_{ki}\log\frac{a_{ki}}{[a^2]_{kj}}.
\end{aligned}
\tag{A11}
$$