A Free Energy Formulation of Music Generation and Perception: Helmholtz Revisited

Karl J. Friston and Dominic A. Friston

1 Introduction

> It is the theory of the sensations of hearing to which the theory of music has to look for the foundation of its structure. (Helmholtz 1877, 4)

This chapter considers music from the point of view of its perception and how acoustic sensations are constructed into musical percepts. Our treatment follows the tradition established by Helmholtz that perception corresponds to inference about the causes of sensations. This therefore requires us to understand the nature of perceptual inference and the formal constraints that this inference places on the nature of music. The basic idea, developed in this chapter, is that music supports the prediction of the unpredictable and that this prediction fulfils a fundamental imperative that we are all compelled to pursue. Heuristically, those activities that we find pleasurable are no more, and no less, than the activities we choose to engage in. The very fact that we can indulge in the same sorts of behaviours repeatedly speaks to the remarkable fact that we are able to maintain a homeostatic exchange with our world—from a physiological to an aesthetic level. We will see later, that this remarkable ability rests upon an active sampling of the sensorium to minimise surprise and fulfil our predictions. In short, music provides the purist opportunity to do what we must do—the opportunity to predict. The opportunity is pure in the sense that musical constructs stand in intimate relation to auditory sensations, perhaps more than any other aesthetic construct:

K. J. Friston (✉)
Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, UK
e-mail: k.friston@ucl.ac.uk

D. A. Friston
Section of Anaesthetics, Pain Medicine and Intensive Care Imperial College London, 369 Fulham Road, London SW10 9NH, UK

R. Bader (ed.), *Sound—Perception—Performance*, Current Research in Systematic Musicology 1, DOI: 10.1007/978-3-319-00107-4_2,

> Music stands in a much closer connection with pure sensation than any other art… In music, the sensations of tone are the material of the art. (Helmholtz 1877 2, 3)

A less heuristic version of this thesis can be motivated from the writings of Helmholtz (1866) on the perceptions in general: in brief, we will consider the brain as a Helmholtz machine (Dayan et al. 1995) that actively constructs predictions or explanations for sensory inputs using internal or generative models. This process of active prediction or inference rests upon predicting the causes of sensory input in a way that minimises prediction errors or surprises. If one generalises this notion of minimising surprise or prediction error to action, one obtains a fairly complete explanation for behaviour as the selective sampling of sensory input to ensure that it conforms to our predictions or expectations, as also articulated nicely by Pearce et al. (2010, 302):

> The ability to anticipate forthcoming events has clear evolutionary advantages, and predictive successes or failures often entail significant psychological and physiological consequences. In music perception, the confirmation and violation of expectations are critical to the communication of emotion and aesthetic effects of a composition.

In what follows, we will see that the ability to predict and anticipate is not just of evolutionary advantage, it is a hallmark of any self-organising biological system that endures in an inconstant and changing world. Active inference then puts prediction centre-stage in the action-perception cycle, to the extent it could be regarded as embodied inference: Embodiment in music production and generation is clearly an important formal constraint on the way music is perceived at many levels. For example, as noted by Helmholtz:

> The enigma which, about 2,500 years ago, Pythagoras proposed to science, which investigates the reasons of things, 'Why is consonance determined by the ratios of small whole numbers?' has been solved by the discovery that the ear resolves all complex sounds into pendular oscillations, according to the laws of sympathetic vibration, and it regards as harmonious only such excitements of nerves as continue without disturbance. (Helmholtz 1877, 279)

In other words, the sensory apparatus and neuronal infrastructure responsible for sensing and predicting auditory input places constraints on what we can predict and, according to the current thesis, what is perceived as musical. This is meant in the sense that colour perception is formally constrained by our (three wavelength selective) photoreceptors to lie in a low dimensional perceptual space—despite the fact that the wavelength composition of visual information arriving at the retina is infinite in its dimensionality. Not only is the perception of music constrained by the embodied brains that perceive it—the nature of music also conforms to embodied constraints on production, so music is "within the compass of executants":

> There is nothing in the nature of music itself to determine the pitch of the tonic of any composition…In short, the pitch of the tonic must be chosen so as to bring the compass of the tones of the piece within the compass of the executants, vocal or instrumental. (Helmholtz 1877, 310)

We will exploit this theme of embodied inference throughout this article; paying careful attention to the neuronal structures that can generate and predict music. This is particularly important for music perception that relies upon neuronal dynamics with deep hierarchical structure.

2 Music and Deep (Hierarchical) Structure

Why is music so compelling to listen to? We started with the premise that music affords the opportunity to predict the unpredictable. This predictable unpredictability rests upon the temporally extensive nature of music and its hierarchical dynamics. If we are biological machines that are built (have evolved) to predict, then the deepest most complicated predictions can only be elicited by stimuli that have a multi-layered (deep) hierarchical structure. Hierarchical causal structure is most evident in a separation of temporal scales, in which slower changes contextualise and prescribe faster changes in a recursive fashion. A non-musical example here would be a story (hours) that unfolds on the basis of a narrative (minutes), which entails semantics that emerge from prosody and syntax (seconds); where the semantics themselves depend upon phonological structures (milliseconds) and so on. Music represents a pure (perhaps the purist) example of deep hierarchical dynamical structure, whose prediction involves the resolution of surprise at multiple temporal scales and—by necessity—can only be accomplished by brains that can support similarly structured neuronal dynamics. We will see an example of this later using the production and recognition of bird songs that are composed by separating the temporal scales of hierarchical neuronal dynamics.

But why the prediction of the unpredictable? It is evident in many writings on music perception and appreciation, that the aesthetic qualities of music and its emotive aspects depend upon a resolution of unpredicted excursions or violations of what might have been predicted. It is the resolution of local violations that is afforded by music's hierarchical structure—and the hierarchical models predicting music. A simple example here would be the use of attractors from dynamical systems theory to produce and recognise musical structures—particularly, attractors that support deterministic chaos. Although this may sound fanciful, these (strange) attractors may play a central role in music and song for several reasons: first, the fact that they exhibit deterministic chaos means that the actual dynamics (say amplitude and frequency modulations as a function of time) are unpredictable from any initial conditions yet, at the same time, they evolve according to entirely deterministic rules which—once inferred by the brain—provide perfect predictions of what will happen next; namely, prediction of the unpredictable. We will see an example of this later, using simulated (bird) songs.

Second, the neuronal dynamics (central pattern generators) responsible for the production of musical stimuli, and—from the perspective of this chapter—their perception can be cast as attractors. Crucially, many of the fundamental aspects of music can be captured quite nicely by attractors with chaotic itinerancy (or related

mathematical images called heteroclinic cycles). This is important because it means that we can simulate or model music perception in a biologically plausible way. Furthermore, by hierarchically composing attractors with different time scales, one can model the perception of auditory objects or scenes with deep hierarchical structure. The last section provides a proof of concept of this approach to music perception, using dynamical attractor models of bird song to simulate both perceptual and neurophysiological responses, of the sort that are seen in real brains.

This chapter comprises four sections: In the first, we briefly review the literature on music and prediction with a special emphasis on the neuroscience of music—as it relates to unconscious inference. The second section introduces an abstract and broad theoretical framework that motivates the importance of prediction and minimising surprise in terms of the free energy principle and active inference. This section is a bit technical but sets up the formalism for the third section that introduces plausible neuronal architectures that minimise surprise (or more exactly maximise free energy) through predictive coding. In the final section, we consider some canonical examples of song perception using simulations of bird song and the predictive coding scheme of the preceding section. These examples illustrate the basic phenomenology and illustrate some ubiquitous phenomena in the neurosciences, like omission responses and categorisation, as measured both psychophysically and electrophysiologically.

3 Prediction in Music and Cognition

> Predictive information processing is fundamental to music in three ways. (1) Prediction and expectancy incorporate the essence of the dynamics of musical temporality. Further they make the experience of local or large-scale goal-directed processes in music possible (based on, e.g., melodic, harmonic or modal features). (2) Predictive processing constitutes a major process involved in musical interaction and synchronisation. (3) Finally, processes of expectancy and prediction are understood to be linked with specific emotional and aesthetic musical effects. (Rohrmeier and Koelsch 2012)

The Helmholtzian view considers the brain as a learning and inference system—assimilating prior beliefs and violated predictions to predict future events as accurately and as parsimoniously as possible. The evolutionary benefits of such a system are clear: it is through a constant updating of our internal model of the world that our interactions with the world are nuanced and optimised. Prediction consequently plays a deep-seated role in all cognition, including that of music. Predictive processing is fundamental to music in three ways (Rohrmeier and Koelsch 2012): it accommodates musical temporality and underlies musical interactions and synchronisation. Furthermore, it plays a key role in mediating the emotive and aesthetic effects of music.

Meyer (1956) proposed that by confirming or violating the listener's musical expectations—and thereby conveying suspense or resolution—music generates an emotional response. However, to recognise the musical qualities of auditory sensations, we must infer the rules or causal structures that underlie our expectations. This structure is manifest over several levels within music, such as melody and harmony, and entails the recognition of rhythmic or metrical structure. Investigations of Meyer's proposal have focused on individual musical features and—consistent with the prominent contribution of harmony to Western styles of music—the processing of harmonic violations has received much attention. Responses reflecting violated predictions, induced by the preceding harmonic context, are measurable with electroencephalography (EEG). For example, infrequent and unpredictable chords, within chord sequences, elicit an early right-anterior negativity (ERAN) and a late bilateral-frontal negativity (N5) in the event related response (ERP) of the listener (Koelsch et al. 2000). These response components are thought to reflect the violation of harmonic expectations and higher processes of harmonic integration into the on-going musical context respectively. Both of their amplitudes are sensitive to the degree of expectance and the probability of harmonic deviation. Further research showed that they are evoked irrespective of whether the stimulus is attended to (Koelsch et al. 2002; Loui et al. 2005). Furthermore, while the N5 is influenced by emotional expression in the performance of a piece (i.e., deliberate variations in loudness and tempo) the ERAN is not (Koelsch et al. 2008). While these components reflect the brain's capacity to establish expectations given a harmonic context, other studies have demonstrated a late positive component with violation of melodic expectations (Besson and Faita 1995; Verleger 1990)—a prediction—dependent response that is modulated by expertise and familiarity.

While the expression of neural responses to violation of musical expectation is established, the implications of these findings for the emotive effects of music are less well understood; due in part to the difficulty of measuring emotional responses. The possibility that emotional attribution is engaged in violation paradigms has emerged as an intriguing prospect from functional neuroimaging studies: unexpected chords activate both the orbital frontolateral cortex (OFLC)—a paralimbic region associated with evaluating a stimulus' emotional salience—and the anterior insula, associated with autonomic responses to emotionally valent stimuli (Koelsch et al. 2005). Interestingly, similar OFLC activation, in subjects listening to classical music, could be prevented when scrambling music, hence disrupting its structure (Levitin and Menon 2003). The authors attributed this activation to the recognition of fine-structured stimuli that evolve over extended periods of time. The absence of this activation, with scrambled music, suggests that high-level (emotive) attributes of music are associated with longer timescales, as suggested by the studies of the (late) N5 ERP components above.

Some studies have used retrospective subjective accounts of emotional and autonomic responses to music, analysing the musical structure of cited excerpts to identify evocative musical characteristics. For example, violations of harmonic

expectation evoke spine shivers, while a musical phrase occurring earlier than expected reliably increases heart rate (Sloboda 1991). While promising, the link between expectation violation and emotion in such approaches is weakened by the lack of an objectively measurable emotional response. Steinbeis et al. (2006) addressed this by combining the use of subjective scales of tension and emotionality with the recording of heart rate and electrodermal activity (EDA)—physiological measures consistently associated with emotional processing. Harmonic expectation was violated via modification of a single chord in each of six Bach chorales. Tension, emotionality and EDA were found to increase with the degree of harmonic expectation violation, which was also reflected in concurrently recorded ERPs (as discussed above).

While the evidence is intriguing, there are some caveats to consider. For example, violation studies are limited by their design, as the violating stimuli do not occur naturally. Furthermore, harmonic studies in particular typically employ paradigms that assess expectation associated with final chords, so the findings may only apply to tonal closure. The generalisation of violation-dependent responses to all aspects of music perception may be more challenging, because some attributes are more predictable than others. For example, while forthcoming elements of melodic or harmonic expectation may be clearly defined, this is not the case for more complex features, such as key structure, that do not necessarily apply to the next musical event, nor to an unambiguous point in time. Additionally, the interplay of attention and prediction in complex styles of music, with non-aligned (e.g. polyrhythmic) features, remains poorly understood (Rohrmeier and Koelsch 2012). Although (unsupervised) statistical learning models can predict single features such as melody (Pearce et al. 2010), predictive models of complex music are still in developmental stages. Nevertheless, the findings thus far support Meyer's proposal, in which music engages the base mechanisms by which we come to understand our environment. The evidence for violation of expectation in emotion discussed here may underlie the initiation of music's intrigue; as Meyer wrote,

> Such states of doubt and confusion are abhorrent. When confronted with them, the mind attempts to resolve them into clarity and certainty. Meyer (1956)

It may indeed be by a flirtatious generation of disorder and its subsequent resolution that music communicates its emotional effect.

3.1 Summary

In summary, the very fact that neuronal responses to the violation of musical predictions can be elicited speaks to the fact that the brain can construct expectations or predictions about the temporal structure of music. Furthermore, the empirical evidence suggests that these predictions have a hierarchical aspect, in which higher level predictions pertain to a longer timescales. Circumstantial evidence suggests that high-level attributes have an emotive dimension; which,

from the point of view of active inference, mean that these representations provide both exteroceptive (auditory) and interoceptive (autonomic) predictions that may explain the visceral responses that music can elicit—visceral responses associated with the violation and resolution of high-level predictions. These conclusions rest upon a hierarchical model of musical structure that is characterised by a separation of temporal scales. In the next section, we will consider the form of hierarchical models that the brain might use; starting with a purely formal or mathematical description that will be used in the subsequent section to understand the neuronal circuits that might underlie hierarchical inference or predictive coding in the brain.

4 Hierarchical Models and Bayesian Inference

In the introduction, we hinted at the fundamental imperative for all self-organising biological systems to minimise their surprise and actively engage with the environment to selectively sample predicted sensations—this is known as active inference. In the previous section, we saw that the brain can predict the underlying causal structure of (musical) sensory streams over multiple timescales; in other words it can infer the hidden causes of its sensory input. In the remainder of this chapter, we try to put these things together and provide a formal model of perception that can be used to illustrate the nature of perception and prediction. In brief, to maintain a homoeostatic exchange with the environment biological systems must counter the dispersion of their sensory states due to environmental fluctuations. In terms of information theory, this means biological systems must minimise the entropy of their sensory states. This can be achieved by minimising the surprise associated with sensory states at each point in time, because (under ergodic assumptions) the long-term average of surprise is entropy. Crucially, negative surprise is the logarithm of Bayesian evidence. This means that minimising surprise is the same as maximising the evidence for a model of the world entailed by the structure and dynamics of the biological system. In other words, we are all obliged to be Bayes-optimal modellers of our sensorium. In what follows, we consider in more detail the nature of the models that the brain may use to guide active inference and, implicitly, make predictions about hidden causes of sensory input.

4.1 Hierarchical Dynamic Models

Hierarchical dynamic models are probabilistic generative models $p(s, \psi) = p(s|\psi)p(\psi)$ based on state-space models. They entail the likelihood $p(s|\psi)$ of getting some sensory data $s(t)$ given some parameters $\psi = \{x, v, \theta\}$ and priors on those parameters $p(\psi)$. We will see that the parameters subsume different quantities, some of which change with time and some which do not. A dynamic model can be written as

$$\begin{aligned} s &= g(x, v) + \omega_s \\ \dot{x} &= f(x, v) + \omega_x \end{aligned} \tag{1}$$

The continuous nonlinear functions (f, g) of the states are parameterized by θ. The states $v(t)$ can be deterministic, stochastic, or both. They are referred to as sources, or causes. The states $x(t)$ mediate the influence of the input on the output and endow the system with memory. They are often referred to as hidden states because they are seldom observed directly. We assume the random fluctuations (i.e., observation noise) $\omega(t)$ are analytic, such that the covariance of the generalised fluctuations $\tilde{\omega} = (\omega, \omega', \omega'', \ldots)$ is well defined. Generalised states include the state itself and all higher order temporal derivatives.

The first (observer) equation above shows that the hidden states (x, v) are needed to generate an output or sensory data. The second (state) equation enforces a coupling between orders of motion of the hidden states and confers memory on the system. Gaussian assumptions about the fluctuations $p(\tilde{\omega}) = \mathcal{N}(0, \Sigma)$ provide the likelihood of any given sensory data and (empirical) priors over the motion of hidden states. It is these empirical priors that can be exploited by the brain to make predictions about the dynamics or trajectories of hidden states causing sensory input. Hierarchical dynamic models have the following form, which generalizes the model in Eq. 1

$$\begin{aligned} s &= g(x^{(1)}, v^{(1)}) + \omega_v^{(1)} \\ \dot{x}^{(1)} &= f(x^{(1)}, v^{(1)}) + \omega_x^{(1)} \\ &\vdots \\ v^{(i-1)} &= g(x^{(i)}, v^{(i)}) + \omega_v^{(i)} \\ \dot{x}^{(i)} &= f(x^{(i)}, v^{(i)}) + \omega_x^{(i)} \\ &\vdots \end{aligned} \tag{2}$$

Again, $f^{(i)} = f(x^{(i)}, v^{(i)})$ and $g^{(i)} = g(x^{(i)}, v^{(i)})$ are continuous nonlinear functions of the states. The random innovations $\omega^{(i)}$ are conditionally independent fluctuations that enter each level of the hierarchy. These play the role of observation error or noise at the first level and induce random fluctuations in the states at higher levels. The causal states $v = (v^{(1)}, v^{(2)}, \ldots)$ link levels, whereas the hidden states $x = (x^{(1)}, x^{(2)}, \ldots)$ link dynamics over time. In hierarchical form, the output of one level acts as an input to the next. Inputs from higher levels can enter nonlinearly into the state equations and can be regarded as changing its control parameters to produce complicated convolutions with "deep" (i.e., hierarchical) structure.

The conditional independence of the fluctuations means that these models have a Markov property over levels (Efron and Morris 1973), which simplifies the architecture of attending inference schemes. See Kass and Steffey (1989) for a discussion of approximate Bayesian inference models of static data and Friston (2008) for dynamic models. In short, a hierarchical form endows models with the

ability to construct their own priors. For example, the prediction $\tilde{g}^{(i)} = \tilde{g}(\tilde{x}^{(i)}, \tilde{v}^{(i)})$ plays the role of a prior expectation on $\tilde{v}^{(i-1)}$, yet it has to be estimated in terms of $(\tilde{x}^{(i)}, \tilde{v}^{(i)})$. This feature is central to many inference and estimation procedures, ranging from mixed-effects analyses in classical covariance component analysis to automatic relevance determination in machine learning.

4.2 Summary

This section has introduced hierarchical dynamic models (in generalized coordinates of motion). These models are about as complicated as one could imagine; they comprise causal and hidden states, whose dynamics can be coupled with arbitrary (analytic) nonlinear functions. Furthermore, these states can have random fluctuations with unknown amplitude and arbitrary (analytic) autocorrelation functions. A key aspect of these models is their hierarchical structure, which induces empirical priors on the causes. These complement the constraints on hidden states, furnished by empirical priors on their motion or dynamics. Later, we will examine the roles of these structural and dynamical priors in perception. We now consider how these models are inverted to disclose the unknown states generating observed sensory data.

4.3 Model Inversion (inference) and Variational Bayes

The concluding part of this section considers model inversion and provides a heuristic summary of the material in Friston (2008). A generative model maps from hidden causes or states to sensory consequences. Recognition (model inversion) inverts this mapping to infer the hidden causes from sensations. We will focus on variational Bayes, which is a generic approach to model inversion that approximates the conditional density $p(\psi|\tilde{s})$ over the unknown states and parameters, given some data. This approximation is achieved by optimizing the sufficient statistics of a recognition density $q(\tilde{\psi})$ over the hidden generalised states, with respect to a lower bound on the log-evidence $\ln p(\tilde{s}|m)$ of the model m (Feynman 1972; Hinton and von Camp 1993; MacKay 1995; Neal and Hinton 1998; Friston 2005; Friston et al. 2006). The log-evidence or negative surprise can be expressed in terms of a free-energy and divergence term

$$\begin{aligned} \ln p(\tilde{s}|m) &= F + D_{KL}(q(\tilde{\psi})||p(\tilde{\psi}|\tilde{s}, m)) \Rightarrow \\ F &= \left\langle \ln p(\tilde{s}, \tilde{\psi}) \right\rangle_q - \left\langle \ln q(\tilde{\psi}) \right\rangle_q \end{aligned} \tag{3}$$

The free-energy comprises an energy term, corresponding to a Gibb's energy, $G(\tilde{s}, \tilde{\psi}) := \ln p(\tilde{s}, \tilde{\psi})$ expected under the recognition density and its entropy. Equation 3 shows that the free energy is a lower-bound on the log-evidence because the divergence term is, by construction, nonnegative. The objective is to optimize the sufficient statistics of the recognition density by maximising the free-energy and minimizing the divergence. This ensures $q(\tilde{\psi}) \approx p(\tilde{\psi}|\tilde{s}, m)$ becomes an approximate posterior density.

Invoking the recognition density converts a difficult integration problem (inherent in computing the evidence) into an easier optimization problem. We now seek a recognition density that maximizes the free energy at each point in time. In what follows, we will assume the hidden parameters are known and focus on the hidden states $u = (x, y)$. To further simplify things, we will assume the brain uses something called the Laplace approximation. This enables us to focus on a single quantity for each unknown state, the conditional expectation or mean. Under the Laplace approximation, the conditional density assumes a fixed Gaussian form $q(\tilde{u}) = \mathcal{N}(\tilde{\mu}, C)$ with sufficient statistics $(\tilde{\mu}, C)$, corresponding to the conditional expectation and covariance of the hidden states. The advantage of the Laplace approximation is that the conditional covariance is a function of the mean (the inverse curvature of the Gibbs energy at the expectation). This means we can reduce model inversion to optimizing one sufficient statistic; namely, the conditional expectation or mean. This is the solution to

$$\dot{\tilde{\mu}} - D\tilde{\mu} = \partial_u F \tag{4}$$

Here, $\dot{\tilde{\mu}} - D\tilde{\mu}$ can be regarded as motion in a frame of reference that moves with the predicted generalised motion $D\tilde{\mu}$, where D is a matrix derivative operator. Critically, the stationary solution (in this moving frame of reference) maximizes free energy. At this point the mean of the motion becomes the motion of the mean, $\dot{\tilde{\mu}} = D\tilde{\mu}$ and $\partial_u F = 0$. Those people familiar with Kalman filtering will see that Eq. 4, can be regarded as (generalised) Bayesian filtering, where the change in conditional expectations $\dot{\tilde{\mu}} = D\tilde{\mu} + \partial_u F$ comprises a prediction and a correction term that depends upon free energy or—as we will see in the next section—prediction error.

4.4 Summary

In this section, we have seen how the inversion of dynamic models can be formulated as an optimization of free energy. By assuming a fixed-form (Laplace) approximation to the conditional density, one can reduce optimization to finding the conditional means of unknown quantities. For the hidden states, this entails finding a path or trajectory that maximizes free energy. This can found by making the motion of the generalized mean perform a gradient ascent in a frame of

reference that moves with the mean of the generalized motion. The only thing we need to implement this recognition scheme (generalised Bayesian filtering) is the Gibbs energy $G(\tilde{s}, \tilde{\psi}) := \ln p(\tilde{s}, \tilde{\psi})$. This is specified completely by the generative model (Eq. 2). In the next section, we look at what this scheme might look like in the brain—and see that it corresponds to something called predictive coding.

5 Hierarchical Models in the Brain

A key architectural principle of the brain is its hierarchical organization (Felleman and van Essen 1991; Maunsell and van Essen 1983; Mesulam 1998; Zeki and Shipp 1988). This has been established most thoroughly in the visual system, where lower (primary) areas receive sensory input and higher areas adopt a multimodal or associational role. The neurobiological notion of a hierarchy rests upon the distinction between forward and backward connections (Angelucci et al. 2002; Felleman and Van Essen 1991; Murphy and Sillito 1987; Rockland and Pandya 1979; Sherman and Guillery 1998). This distinction is based upon the specificity of cortical layers that are the predominant sources and origins of extrinsic connections. Forward connections arise largely in superficial pyramidal cells, in supra-granular layers, and terminate on spiny stellate cells of layer four in higher cortical areas (DeFelipe et al. 2002; Felleman and Van Essen 1991). Conversely, backward connections arise largely from deep pyramidal cells in infra-granular layers and target cells in the infra and supra-granular layers of lower cortical areas. Intrinsic connections mediate lateral interactions between neurons that are a few millimetres away. There is a key functional asymmetry between forward and backward connections that renders backward connections more modulatory or nonlinear in their effects on neuronal responses (e.g., Sherman and Guillery 1998; see also Hupe et al. 1998). This is consistent with the deployment of voltage-sensitive NMDA receptors in the supra-granular layers that are targeted by backward connections (Rosier et al. 1993). Typically, the synaptic dynamics of backward connections have slower time constants. This has led to the notion that forward connections are driving and elicit obligatory responses in higher levels, whereas backward connections have both driving and modulatory effects and operate over larger spatial and temporal scales.

5.1 Bayesian Filtering and Predictive Coding

This hierarchical structure of the brain speaks to hierarchical models of sensory input. We now consider how the brain‘s functional architecture can be understood as inverting hierarchical models (recovering the hidden causes of sensory input). If we assume that the activity of neurons encodes the conditional mean of states, then Eq. 4 specifies the neuronal dynamics entailed by perception or recognizing states

of the world from sensory data. In Friston (2008), we show how these dynamics can be expressed simply in terms of prediction errors on the causes and motion of the hidden states. Using these errors, we can write Eq. 4 as

$$\begin{aligned}\dot{\tilde{\mu}}_v^{(i)} &= \mathcal{D}\tilde{\mu}_v^{(i)} - \partial_{\tilde{v}}\tilde{\varepsilon}^{(i)} \cdot \xi^{(i)} - \xi_v^{(i+1)} \\ \dot{\tilde{\mu}}_x^{(i)} &= \mathcal{D}\tilde{\mu}_x^{(i)} - \partial_{\tilde{x}}\tilde{\varepsilon}^{(i)} \cdot \xi^{(i)} \\ \\ \xi_v^{(i)} &= \Pi_v^{(i)}(\tilde{\mu}_v^{(i-1)} - g^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)})) \\ \xi_x^{(i)} &= \Pi_x^{(i)}(\mathcal{D}\tilde{\mu}_x^{(i)} - f^{(i)}(\tilde{\mu}_x^{(i)}, \tilde{\mu}_v^{(i)}))\end{aligned} \tag{5}$$

This scheme describes a gradient descent on the (sum of squared) prediction error—or more exactly precision-weighted prediction errors $\xi^{(i)} = \Pi^{(i)}\tilde{\varepsilon}^{(i)}$, where precision $\Pi^{(i)}$ corresponds to the reliability (inverse covariance) of the prediction error $\tilde{\varepsilon}^{(i)}$ at the i-th level of the hierarchy. The first pair of equalities just says that conditional expectations about hidden causes and states $(\tilde{\mu}_v^{(i)}, \tilde{\mu}_x^{(i)})$ are updated based upon the way we would predict them to change—the first term—and subsequent terms that minimise prediction error. The second pair expresses prediction error as the conditional expectations about hidden causes and (the changes in) hidden states minus their predicted values.

It is difficult to overstate the generality and importance of Eq. 5—it grandfathers nearly every known statistical estimation scheme, under parametric assumptions about additive noise. These range from ordinary least squares to advanced Bayesian filtering schemes (see Friston 2008). In this general setting, Eq. 5 corresponds to (generalised) predictive coding. Under linear models, it reduces to linear predictive coding, also known as Kalman-Bucy filtering.

5.2 *Predictive Coding and Message Passing in the Brain*

In neuronal network terms, Eq. 5 says that prediction error units receive messages based on expectations in the same level and the level above. This is because the hierarchical form of the model only requires expectations between neighbouring levels to form prediction errors. Conversely, expectations are driven by prediction error in the same level and the level below—updating expectations about hidden states and causes respectively. This updating corresponds to an accumulation of prediction errors—in that the rate of change of conditional expectations is proportional to prediction error. This means expectations are proportional to the integral or accumulation of prediction errors over time. Crucially, this accumulation requires only the prediction error from the lower level and the level in question. These constitute the bottom-up and lateral messages that drive conditional expectations to provide better predictions—or representations—which suppress the prediction error. Electrophysiologically, this means that one would

expect to see a transient prediction error response to bottom-up afferents (in neuronal populations encoding prediction error) that is suppressed to baseline firing rates by sustained responses (in neuronal populations encoding predictions). This is the essence of recurrent message passing between hierarchical levels to suppress prediction error (see Fig. 1 and Friston 2008 for a more detailed discussion).

We can identify error-units with superficial pyramidal cells, because the only messages that pass up the hierarchy are prediction errors and superficial pyramidal cells originate forward connections in the brain. This is useful because it is these cells that are primarily responsible for electroencephalographic (EEG) signals that can be measured noninvasively. Similarly, the only messages that are passed down the hierarchy are the predictions from state-units that are necessary to form

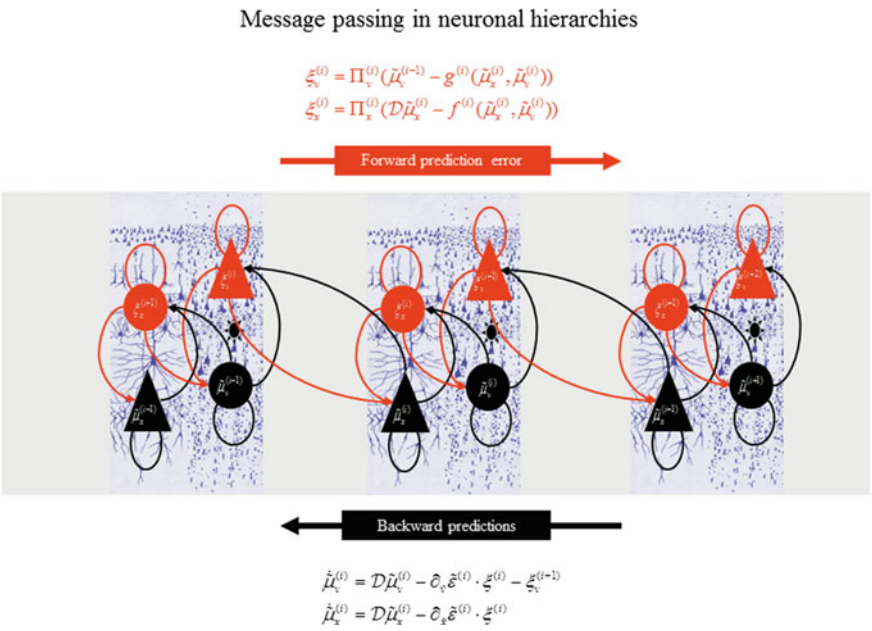


Fig. 1 Schematic, detailing the neuronal architectures that encode a recognition density over the hidden states of a hierarchical model. This schematic shows the speculative cells of origin of forward (*driving*) connections that convey prediction error from a lower area to a higher area and the backward connections that are used to construct predictions. These predictions try to explain away input from lower areas by suppressing prediction error. In this scheme, the sources of forward connections are the superficial pyramidal cell population, and the sources of backward connections are the deep pyramidal cell population. The differential equations relate to the optimization scheme detailed in the main text. The state-units and their efferents are in black and the error-units in red, with causes on the right and hidden states on the left. For simplicity, we have assumed the output of each level is a function of, and only of, the hidden states. This induces a hierarchy over levels and, within each level, a hierarchical relationship between states, where causes predict hidden states. This schematic shows how the neuronal populations may be deployed hierarchically within three cortical areas (or *macro-columns*)

prediction errors in lower levels. The sources of extrinsic backward connections are the deep pyramidal cells, and one might deduce that these encode the expected causes of sensory states (see Mumford 1992 and Fig. 1). Crucially, the motion of each state-unit is a linear mixture of bottom-up prediction error (Eq. 5). This is exactly what is observed physiologically, in that bottom-up driving inputs elicit obligatory responses that do not depend on other bottom-up inputs. The prediction error itself is formed by predictions conveyed by backward and lateral connections. These influences embody the nonlinearities implicit in $\tilde{g}^{(i)}$ and $\tilde{f}^{(i)}$. Again, this is entirely consistent with the nonlinear or modulatory characteristics of backward connections.

5.3 Summary

In summary, we have seen how the inversion of a generic hierarchical and dynamical model of sensory inputs can be transcribed onto neuronal quantities that optimize a variational free energy bound on the evidence for that model. This optimization corresponds, under some simplifying assumptions, to suppression of prediction error at all levels in a cortical hierarchy. This suppression rests upon a balance between bottom-up (prediction error) influences and top-down (empirical prior) influences. In the final section, we use this scheme to simulate neuronal responses. Specifically, we pursue the electrophysiological correlates of prediction error and ask whether we can understand the violation phenomena in event-related potential (ERP), discussed in Sect. 2, in terms of hierarchical inference and message passing in the brain.

6 Birdsong and Attractors

In this section, we examine the emergent properties of a system that uses hierarchical dynamics or attractors as generative models of sensory input. The example we use is birdsong, and the empirical measures we focus on are local field potentials (LFPs) or evoked (ERP) responses that can be recorded noninvasively. Our aim is to show that canonical features of empirical electrophysiological responses can be reproduced easily under attractor models of sensory input. Furthermore, in a hierarchical setting, the use of dynamic models has some interesting implications for perceptual infrastructures (Kiebel et al. 2008). The examples in this section are taken from Friston and Kiebel (2009), to which the reader is referred for more details.

We first describe the model of birdsong and demonstrate the nature and form of this model through simulated lesion experiments. This model is then used to reproduce the violation-dependent responses discussed in Sect. 2 using, perhaps, the most profound form of violation; namely, an omission of an expected event.

We will then use simplified versions of this model to show how attractors can be used to categorize sequences of stimuli quickly and efficiently. Throughout this section, we will exploit the fact that superficial pyramidal cells are the major contributors to observed LFP and ERP signals, which means we can ascribe these signals to prediction error because the superficial pyramidal cells are the source of bottom-up messages in the brain (see Fig. 1).

6.1 Attractors in the Brain

The basic idea in this chapter is that the environment unfolds as an ordered sequence of spatiotemporal dynamics, whose equations of motion entail attractor manifolds that contain sensory trajectories. Critically, the shape of the manifold generating sensory data is itself changed by other dynamical systems that could have their own attractors. If we consider the brain has a generative model of these coupled dynamical systems, then we would expect to see attractors in neuronal dynamics that are trying to predict sensory input. In a hierarchical setting, the states of a high-level attractor enter the equations of motion of a low-level attractor in a nonlinear way, to change the shape of its manifold. This form of generative model has a number of sensible and appealing characteristics.

First, at any level the model can generate and therefore encode structured sequences of events, as the states flow over different parts of the manifold. These sequences can be simple, such as the quasi-periodic attractors of central pattern generators (McCrea and Rybak 2008) or can exhibit complicated sequences of the sort associated with chaotic and itinerant dynamics (e.g., Breakspear and Stam 2005; Canolty et al. 2006; Friston 1997; Haken Kelso et al. 1990; Jirsa et al. 1998; Kopell et al. 2000; Rabinovich et al. 2008).

Second, hierarchically deployed attractors enable the brain to generate and therefore predict or represent different categories of sequences. This is because any low-level attractor embodies a family of trajectories that correspond to a structured sequence. The neuronal activity encoding the particular state at any one time determines where the current dynamics are within the sequence, while the shape of the attractor manifold determines which sequence is currently being expressed. In other words, the attractor manifold encodes what is being perceived and the neuronal activity encodes where the current percept is located on the manifold or within the sequence.

Third, if the state of a higher attractor changes the manifold of a subordinate attractor, then the states of the higher attractor come to encode the category of the sequence or dynamics represented by the lower attractor. This means it is possible to generate and represent sequences of sequences and, by induction, sequences of sequences of sequences, and so on. This rests upon the states of neuronal attractors at any cortical level providing control parameters for attractor dynamics at the level below. This necessarily entails a nonlinear interaction between the top-down

effects of the higher attractor and the states of the recipient attractor. Again, this is entirely consistent with the known functional asymmetries between forward and backward connections and speaks to the nonlinear effects of top-down connections in the real brain.

Finally, this particular model has implications for the temporal structure of perception and, in particular, music. Put simply, the dynamics of high-level representations unfold more slowly than the dynamics of lower level representations. This is because the state of a higher attractor prescribes a manifold that guides the flow of lower states. In the limiting case of the higher level having a fixed-point attractor, its fixed states will encode lower level dynamics, which could change quite rapidly. We will see an example of this later, when considering the perceptual categorization of different sequences of chirps subtending birdsongs. This attribute of hierarchically coupled attractors enables the representation of arbitrarily long sequences of sequences and suggests that neuronal representations in the brain will change more slowly at higher levels (Kiebel et al. 2008; see also Botvinick et al. 2007; Hasson et al. 2008). One can turn this argument on its head and use the fact that we are able to recognize sequences of sequences (e.g., Chait et al. 2007) as an existence proof for this sort of generative model. In the examples that follow, we will try to show how autonomous dynamics furnish generative models of sensory input, which behave much like real brains, when measured electrophysiologically.

6.2 A Synthetic Avian Brain

The toy example used here deals with the generation and recognition of birdsongs (Laje and Mindlin 2002). We imagine that birdsongs are produced by two time-varying control parameters that control the frequency and amplitude of vibrations emanating from the syrinx of a songbird (see Fig. 2). There has been an extensive modelling effort using attractor models at the biomechanical level to understand the generation of birdsong (e.g., Laje et al. 2002). Here, we use the attractors at a higher level to provide time-varying control over the resulting sonograms. We drive the syrinx with two states of a Lorenz attractor, one controlling the frequency (between 2 and 5 kHz) and the other (after rectification) controlling the amplitude or volume. The parameters of the Lorenz attractor were chosen to generate a short sequence of chirps every second or so. To endow the generative model with a hierarchical structure, we placed a second Lorenz attractor, whose dynamics were an order of magnitude slower, over the first. The states of the slower attractor entered as control parameters (known as the Rayleigh and Prandtl number) to control the dynamics exhibited by the first. These dynamics could range from a fixed-point attractor, where the states of the first are all zero, through to quasi-periodic and chaotic behaviour, when the value of the Prandtl number exceeds an appropriate threshold (about 24) and induces a bifurcation. Because higher states

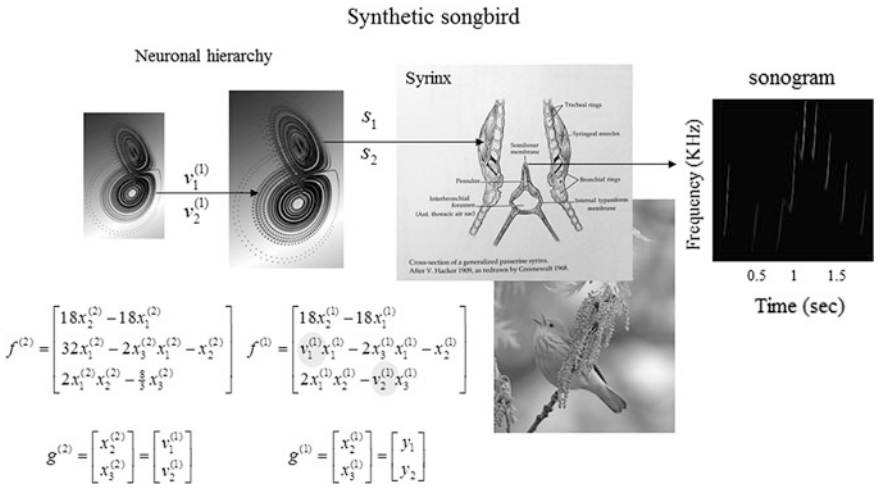


Fig. 2 Schematic, showing the construction of a generative model for birdsongs. This model comprises two Lorenz attractors, where the higher attractor delivers two control parameters (*grey circles*) to a lower level attractor, which, in turn, delivers two control parameters to a synthetic syrinx to produce amplitude and frequency modulated stimuli. This stimulus is represented as a sonogram in the right panel. The equations represent the hierarchical dynamic model in the form of Eq. 2

evolve more slowly, they switch the lower attractor on and off, generating distinct songs, where each song comprises a series of distinct chirps (see Fig. 3).

6.3 Song Recognition

This model generates spontaneous sequences of songs using autonomous dynamics. We generated a single song, corresponding roughly to a cycle of the higher attractor and then inverted the ensuing sonogram (summarized as peak amplitude and volume) using the message-passing scheme described in previous sections. The results are shown in Fig. 3 and demonstrate that, after several hundred milliseconds, the veridical hidden states and superordinate causes can be recovered. Interestingly, the third chirp is not perceived, in that the first-level prediction error was not sufficient to overcome the dynamical and structural priors entailed by the model. However, once the subsequent chirp had been predicted correctly the following sequence of chirps was recognized with a high degree of conditional confidence. Note that when the second and third chirps in the sequence are not recognized, first-level prediction error is high and the conditional confidence about the causes at the second level is low (reflected in the wide 90 % confidence intervals). Heuristically, this means that the synthetic bird listening to

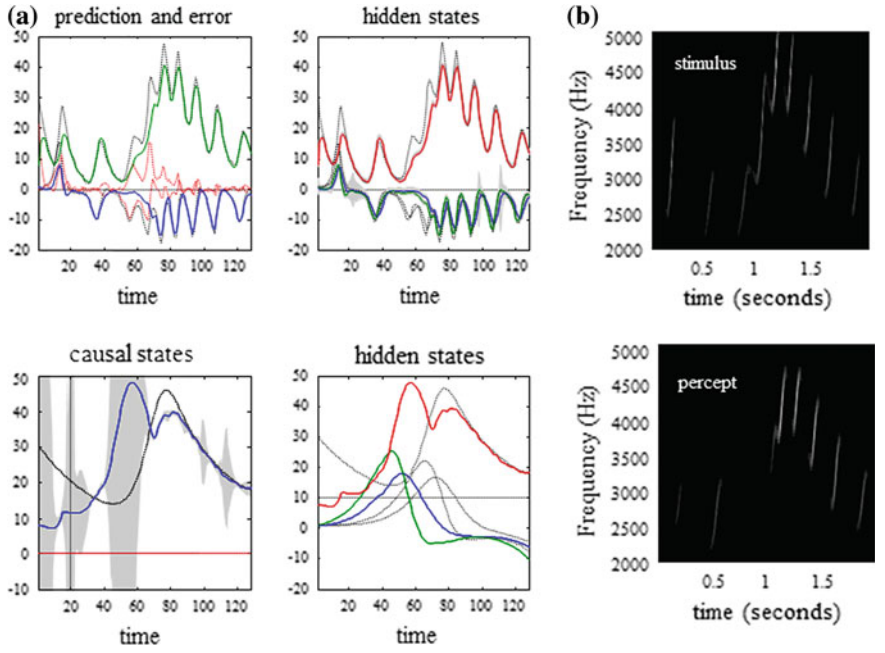


Fig. 3 Results of a Bayesian inversion or deconvolution of the sonogram shown in Fig. 2. **a** Upper panels show the time courses of hidden and causal states. (*Upper left*) These are the true and predicted states driving the syrinx and are simple mappings from two of the three hidden states of the first-level attractor. The solid lines respond to the conditional mode and the dotted lines to the true values. The discrepancy is the prediction error and is shown as a broken red line. (*Upper right*) The true and estimated hidden states of the first-level attractor. Note that the third hidden state has to be inferred from the sensory data. Confidence intervals on the conditional expectations are shown in grey and demonstrate a high degree of confidence, because a low level of sensory noise was used in these simulations. The panels below show the corresponding causes and hidden states at the second level. Again the conditional expectations are shown as solid lines and the true values as broken lines. Note the inflated conditional confidence interval halfway through the song when the third and fourth chirps are misperceived. **b** The stimulus and percept in sonogram format, detailing the expression of different frequencies generated over peristimulus time

the song did not know which song was being emitted and was unable to predict subsequent chirps.

6.4 Structural and Dynamic Priors

This example provides a nice opportunity to illustrate the relative roles of structural and dynamic priors. Structural priors are provided by the top-down inputs that dynamically reshape the manifold of the low-level attractor. However, this

attractor itself contains an abundance of dynamical priors that unfold in generalized coordinates. Both provide important constraints on the evolution of sensory states, which facilitate recognition. We can selectively destroy these priors by lesioning the top-down connections to remove structural priors (over hidden causes) or by cutting the intrinsic connections that mediate dynamic priors (over hidden states). The latter involves cutting the self-connections in Fig. 1 among the causal and state units. The results of these two simulated lesion experiments are shown in Fig. 4. The top panel shows the percept as in the previous panel, in terms of the predicted sonogram and prediction error at the first and second level. The subsequent two panels show exactly the same information but without structural (middle) and dynamic (lower) priors. In both cases, the bird fails to recognize the sequence with a corresponding inflation of prediction error, particularly at the last level. Interestingly, the removal of structural priors has a less marked effect on recognition than removing the dynamical priors. Without dynamical priors there is a failure to segment the sensory stream and although there is a preservation of frequency tracking, the dynamics *per se* have completely lost their sequential structure. Although it is interesting to compare structural and dynamics priors, the important message here is that both are necessary for veridical perception and that removal of either leads to suboptimal inference. Both of these empirical priors prescribe dynamics that enable the synthetic bird to predict what will be heard next. This leads to the question 'What would happen if the song terminated prematurely?'

6.5 Omission and Violation of Predictions

We repeated the above simulation but terminated the song after the fifth chirp. The corresponding sonograms and percepts are shown with their prediction errors in Fig. 5. The left panels show the stimulus and percept as in Fig. 4, while the right panels show the stimulus and responses to omission of the last syllables. These results illustrate two important phenomena. First, there is a vigorous expression of prediction error after the song terminates abruptly. This reflects the dynamical nature of the recognition process because, at this point, there is no sensory input to predict. In other words, the prediction error is generated entirely by the predictions afforded by the dynamic model of sensory input. It can be seen that this prediction error (with a percept but no stimulus) is almost as large as the prediction error associated with the third and fourth stimuli that are not perceived (stimulus but no percept). Second, it can be seen that there is a transient percept, when the omitted chirp should have occurred. Its frequency is slightly too low, but its timing is preserved in relation to the expected stimulus train. This is an interesting stimulation from the point of view of ERP studies of omission-related responses. These simulations and related empirical studies (e.g., Nordby et al. 1994; Yabe et al. 1997) provide clear evidence for the predictive capacity of the brain. In the context

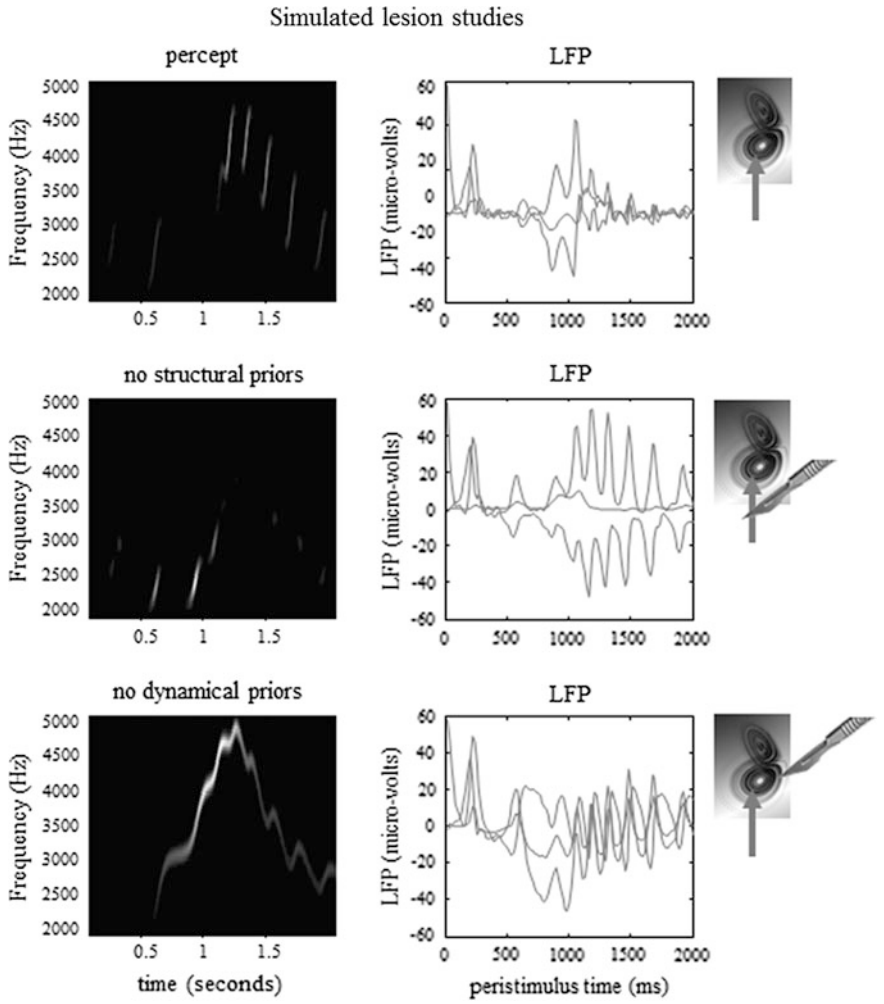
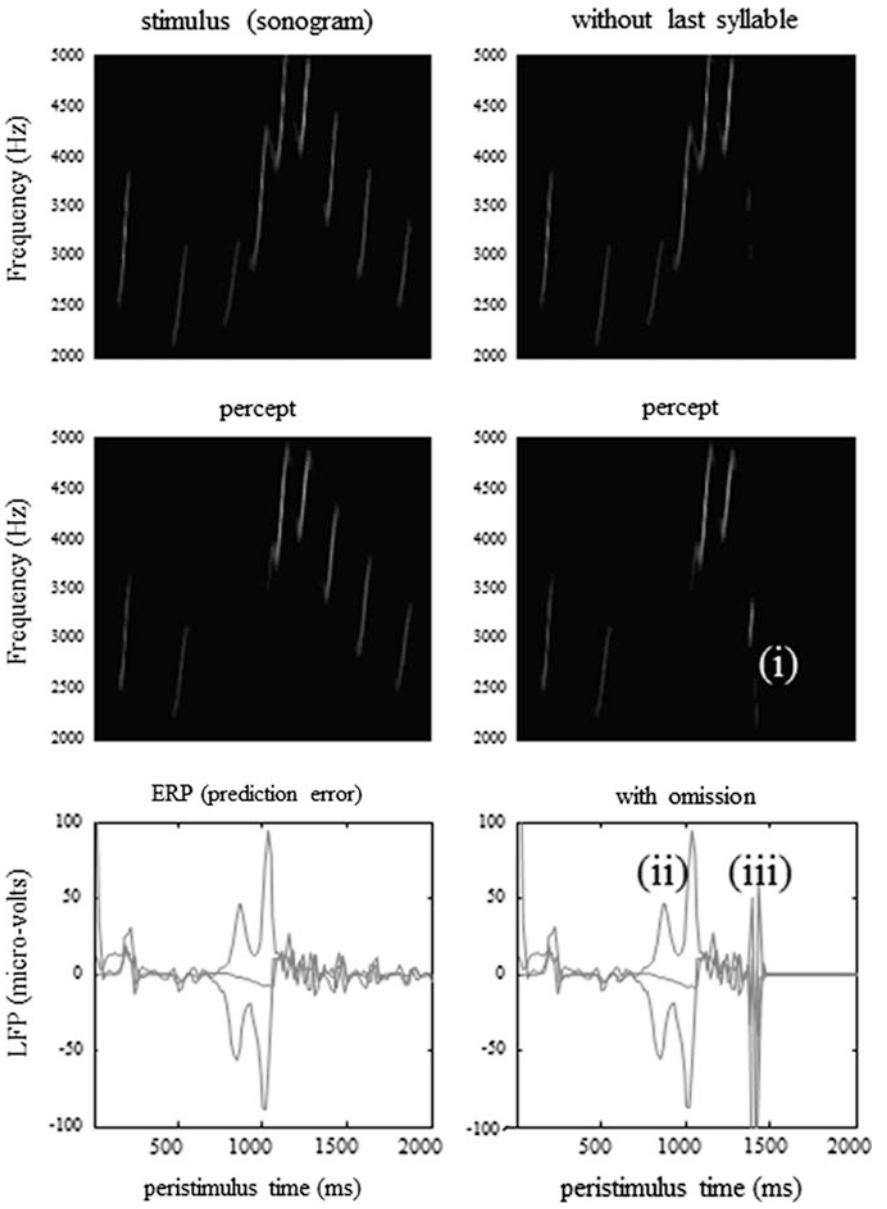


Fig. 4 Results of simulated lesion studies using the birdsong model of the previous figures. The left panels show the percept in terms of the predicted sonograms, and the right panels show the corresponding prediction error (at both levels); these are the differences between the incoming sensory information and the prediction and the discrepancy between the conditional expectation of the second-level cause and that predicted by the second-level hidden states. *Top panels* The recognition dynamics in the intact bird. *Middle panels* The percept and corresponding prediction errors when the connections between the hidden states at the second level and their corresponding causes are removed. This effectively removes structural priors on the evolution of the attractor manifold prescribing the sensory dynamics at the first level. *Lower panels* The effects of retaining the structural priors but removing the dynamical priors by cutting the connections that mediate inversion in generalized coordinates. These results suggest that both structural and dynamical priors are necessary for veridical perception

of music studies, the results in Fig. 5 can be seen as a rough model of the violation (harmonic) responses described in Sect. 2 (Koelsch et al. 2000, 2002, 2008, Loui et al. 2005; Besson and Faita 1995; Verleger 1990). In this example, prediction rests upon the internal construction of an attractor manifold that defines a family of trajectories, each corresponding to the realization of a particular song. In the last simulation we look more closely at perceptual categorization of these songs.

6.6 Perceptual Categorization

In the previous simulations, we saw that a song corresponds to a sequence of chirps that is preordained by the shape of an attractor manifold controlled by top-down inputs. This means that for every point in the state-space of the higher attractor there is a corresponding manifold or category of song. In other words, recognizing or categorizing a particular song corresponds to finding a fixed location in the higher state-space. This provides a nice metaphor for perceptual categorization; because the neuronal states of the higher attractor represent, implicitly, a category of song. Inverting the generative model means that, probabilistically, we can map from a sequence of sensory events to a point in some perceptual space, where this mapping corresponds to perceptual recognition or categorization. This can be demonstrated in our synthetic songbird by ignoring the dynamics of the second-level attractor and exposing the bird to a song and letting the states at the second level optimize their location in perceptual space, to best predict the sensory input. To illustrate this, we generated three songs by fixing the Rayleigh and Prandtl variables to three distinct values. We then placed uninformative priors on the second-level causes (that were previously driven by the hidden states of the second-level attractor) and inverted the model in the usual way. Figure 6a shows the results of this simulation for a single song. This song comprises a series of relatively low-frequency chirps emitted every 250 ms or so. The causes of this song (song C in panel b) are recovered after the second chirp, with relatively tight confidence intervals (the blue and green lines in the lower left panel). We then repeated this exercise for three songs. The results are shown in Fig. 6b. The songs are portrayed in sonogram format in the top panels and the inferred perceptual causes in the bottom panels. The left panel shows the evolution of these causes for all three songs as a function of peristimulus time and the right shows the corresponding conditional density in the causal or perceptual space of these two states after convergence. It can be seen that for all three songs the 90 % confidence interval encompasses the true values (red dots). Furthermore, there is very little overlap between the conditional densities (grey regions), which means that the precision of the perceptual categorization is almost 100 %. This is a simple but nice example of perceptual categorization, where sequences of sensory events


stimulus (sonogram)
without last syllable
Frequency (Hz)
5000
4500
4000
3500
3000
2500
2000
percept
percept
(i)
ERP (prediction error)
with omission
LFP (micro-volts)
100
50
0
-50
-100
(ii)
(iii)
500
1000
1500
2000
peristimulus time (ms)
peristimulus time (ms)

◀ **Fig. 5** Omission-related responses. Here, we have omitted the last few chirps from the stimulus. The left-hand panels show the original sequence and responses evoked. The right-hand panels show the equivalent dynamics on omission of the last chirps. The top panels show the stimulus and the middle panels the corresponding percept in sonogram format. The interesting thing to note here is the occurrence of an anomalous percept after termination of the song on the lower right (*i*). This corresponds roughly to the chirp that would have been perceived in the absence of omission. The lower panels show the corresponding (precision-weighted) prediction error under the two stimuli at both levels. A comparison of the two reveals a burst of prediction error when a stimulus is missed (*ii*) and at the point that the stimulus terminates (*iii*) despite the fact that there is no stimulus present at this time. The darker lines correspond to prediction error at the first level, and the lighter lines correspond to prediction error at the second level

with extended temporal support can be mapped to locations in perceptual space, through Bayesian filtering (predictive coding) of the sort entailed by the free-energy principle.

7 Conclusion

This chapter has suggested that the architecture of cortical systems speaks to hierarchical generative models in the brain. The estimation or inversion of these models corresponds to a generalized Bayesian filtering (predictive coding) of sensory inputs to disclose their causes. This predictive coding can be implemented in a neuronally plausible fashion, where neuronal dynamics self-organize when exposed to inputs to suppress prediction errors. The focus of this chapter has been on the nature of the hierarchical models and, in particular, models that show autonomous dynamics. These models may be relevant for music perception because they enable sequences of sequences to be inferred or recognized. We have tried to demonstrate their plausibility, in relation to empirical observations, by interpreting the prediction error, associated with model inversion, with observed electrophysiological responses. These models provide a graceful way to map from complicated sensory trajectories to points in abstract perceptual spaces. Furthermore, in a hierarchical setting, this mapping may involve trajectories in perceptual spaces of increasingly higher order. The mathematical formalism (and simulations) of hierarchical Bayesian inference in the brain provides a nice link between the generic principles of perceptual inference (and self-organisation) and the perception of music—in particular, it enabled us to simulate one of the most prominent electrophysiological phenomena in music research; namely violation—dependent responses.

The ideas presented in this chapter have a long history, starting with the notion of neuronal energy (Helmholtz 1860), covering ideas like efficient coding and analysis by synthesis (Barlow 1961; Neisser 1967) to more recent formulations in terms of Bayesian inversion and predictive coding (e.g., Ballard et al.1983; Dayan et al. 1995; Kawato et al. 1993; Mumford 1992; Rao and Ballard 1998). This work

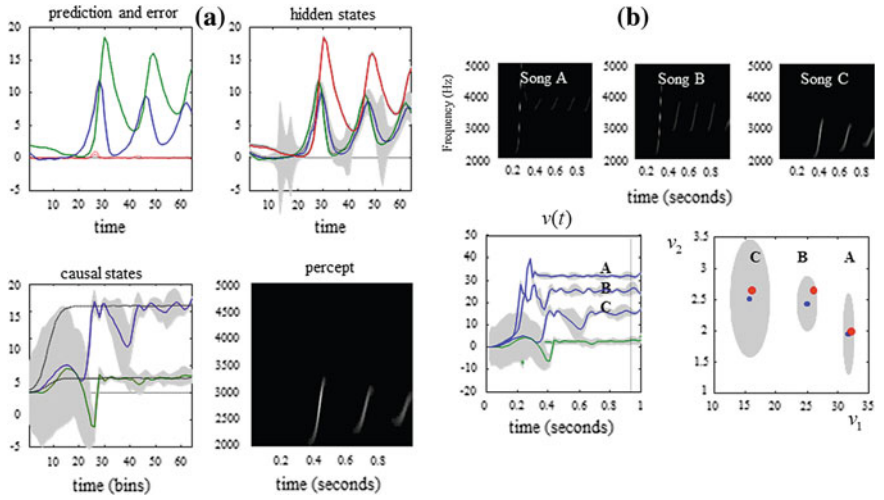


Fig. 6 **a** Schematic demonstration of perceptual categorization. This figure follows the same format as Fig. 3. However, here there are no hidden states at the second level, and the causes were subject to stationary and uninformative priors. This song was generated by a first-level attractor with fixed control parameters of 16 and 8/3, respectively. It can be seen that, on inversion of this model, these two control variables, corresponding to causes or states at the second level, are recovered with relatively high conditional precision. However, it takes about 50 iterations (about 600 ms) before they stabilize. In other words, the sensory sequence has been mapped correctly to a point in perceptual space after the occurrence of the second chirp. This song corresponds to song C on the right. **b** The results of inversion for three songs each produced with three distinct pairs of values for the second-level causes (the Rayleigh and Prandtl variables of the first-level attractor). *Upper panel* The three songs shown in sonogram format corresponding to a series of relatively high-frequency chirps that fall progressively in both frequency and number as the Rayleigh number is decreased. *Lower left* These are the second-level causes shown as a function of peristimulus time for the three songs. It can be seen that the causes are identified after about 600 ms with high conditional precision. *Lower right* This shows the conditional density on the causes shortly before the end of peristimulus time (*dotted line on the left*). The blue dots correspond to conditional means or expectations, and the grey areas correspond to the conditional confidence regions. Note that these encompass the true values (*red dots*) used to generate the songs. These results indicate that there has been a successful categorization, in the sense that there is no ambiguity (from the point of view of the synthetic bird) about which song was heard

has tried to provide support for the notion that the brain uses attractors to represent and predict causes in the sensorium (Byrne et al. 2007; Deco and Rolls 2003; Freeman 1987; Tsodyks 1999). More generally, one might conclude that we need large brains, with deep hierarchical structure, to perceive and appreciate the world we inhabit. This resonates with Einstein's conclusion:

> He who joyfully marches to music in rank and file has already earned my contempt. He has been given a large brain by mistake, since for him the spinal cord would suffice. Albert Einstein

although motivated from a slightly different perspective.

Acknowledgments The Wellcome Trust funded this work. We would also like to thank Larry Goodyer for invaluable discussions.

References

Angelucci, A., Levitt, J. B., Walton, E. J., Hupe, J. M., Bullier, J., & Lund, J. S. (2002). Circuits for local and global signal integration in primary visual cortex. *Journal of Neuroscience, 22*, 8633–8646.

Ballard, D. H., Hinton, G. E., & Sejnowski, T. J. (1983). Parallel visual computation. *Nature, 306*, 21–26.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Besson, M., & Faita, F. (1995). Event-related potential (ERP) study of musical expectancy—comparison of musicians with nonmusicians. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1278–1296.

Botvinick, M. M. (2007). Multilevel structure in behaviour and in the brain: A model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1485), 1615–1626.

Breakspear, M., & Stam, C. J. (2005). Dynamics of a neural system with a multiscale architecture. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*, 1051–1107.

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychology Review, 114*(2), 340–375.

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., et al. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science, 313*, 1626–1628.

Chait, M., Poeppel, D., de Cheveigné, A., & Simon, J. Z. (2007). Processing asymmetry of transitions between order and disorder in human auditory cortex. *Journal of Neuroscience, 27*(19), 5207–5514.

Dayan, P., Hinton, G. E., & Neal, R. M. (1995). The Helmholtz machine. *Neural Computation, 7*, 889–904.

Deco, G., & Rolls, E. T. (2003). Attention and working memory: A dynamical model of neuronal activity in the prefrontal cortex. *European Journal of Neuroscience, 18*(8), 2374–2390.

DeFelipe, J., Alonso-Nanclares, L., & Arellano, J. I. (2002). Microstructure of the neocortex: Comparative aspects. *Journal of Neurocytology, 31*, 299–316.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association, 68*, 117–130.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*, 1–47.

Feynman, R. P. (1972). *Statistical mechanics*. Reading, MA: Benjamin.

Freeman, W. J. (1987). Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. *Biological Cybernetics, 56*(2–3), 139–150.

Friston, K. J . (1997). Transients, metastability, and neuronal dynamics. *NeuroImage, 5*(2), 164–171.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360*, 815–836.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris, 100*(1–3), 70–87.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11), e1000211.

Friston, K., & Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks, 22*(8), 1093–1104.

Haken, H., Kelso, J. A. S., Fuchs, A., & Pandya, A. S. (1990). Dynamic pattern-recognition of coordinated biological motion. *Neural Networks, 3*, 395–401.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience, 28*, 2539–2550.

Helmholtz, H. (1860/1962). Handbuch der physiologischen Optik. In J. P. C. Southall (Ed.) (Vol. 3). New York: Dover, (Translation).

Helmholtz, H. (1866/1962). Concerning the perceptions in general. In J. Southall (Ed.) *Treatise on physiological optics* (3rd ed., Vol. III). New York: Dover, (Translation).

Helmholtz, H. (1877). On the sensations of tone as a physiological basis for the theory of music. In A. J. Ellis (Ed.), *Fourth German edition, translated, revised, corrected with notes and additional appendix*. New York: Dover Publications Inc., 1954 (Reprint).

Hinton, G. E., & von Camp, D. (1993). Keeping neural networks simple by minimising the description length of weights. *In Proceedings of COLT-93*, 5–13.

Hupe, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature, 394*, 784–787.

Jirsa, V. K., Fuchs, A., & Kelso, J. A. (1998). Connecting cortical and behavioral dynamics: bimanual coordination. *Neural Computation, 10*, 2019–2045.

Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association, 407*, 717–726.

Kawato, M., Hayakawa, H., & Inui, T. (1993). A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network, 4*, 415–422.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4*(11), e1000209.

Koelsch, S., Gunter, T., Friederici, A. D., & Schröger, E. (2000). Brain indices of music processing: "nonmusicians" are musical. *Journal of Cognitive Neuroscience, 12*(3), 520–541.

Koelsch, S., Schroger, E., & Gunter, T. C. (2002). Music matters: Preattentive musicality of the human brain. *Psychophysiology, 39*(1), 38–48.

Koelsch, S., Fritz, T., Schulze, K., Alsop, D., & Schlaug, G. (2005). Adults and children processing music: An fMRI study. *Neuroimage, 25*(4), 1068–1076.

Koelsch, S., Kilches, S., Steinbeis, N., & Schelinski, S. (2008). Effects of unexpected chords and of performer's expression on brain responses and electrodermal activity. *PloS ONE, 3*, e2631.

Kopell, N., Ermentrout, G. B., Whittington, M. A., & Traub, R. D. (2000). Gamma rhythms and beta rhythms have different synchronization properties. *Proceedings of the National Academy of Sciences USA, 97*, 1867–1872.

Laje, R., Gardner, T. J., & Mindlin, G. B. (2002). Neuromuscular control of vocalizations in birdsong: a model. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 65, 051921.1–8.

Laje, R., & Mindlin, G. B. (2002). Diversity within a birdsong. *Physical Review Letters, 89*, 288102.

Levitin, D. J., & Menon, V. (2003). Musical structure is processed in "language" areas of the brain: a possible role for Brodmann Area 47 in temporal coherence. *Neuroimage, 20*(4), 2142–2152.

Loui, P., Grent-'t-Jong, T., Torpey, D., & Woldorff, M. (2005). Effects of attention on the neural processing of harmonic syntax in Western music. *Brain Research. Cognitive Brain Research, 25*(3), 678–687.

MacKay, D. J. C. (1995). Free-energy minimisation algorithm for decoding and cryptoanalysis. *Electronics Letters, 31*, 445–447.

Maunsell, J. H., & van Essen, D. C. (1983). The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience, 3*, 2563–2586.

McCrea, D. A., & Rybak, I. A. (2008). Organization of mammalian locomotor rhythm and pattern generation. *Brain Research Reviews, 57*(1), 134–146.
Mesulam, M. M. (1998). From sensation to cognition. *Brain*, *121*, 1013–1052.
Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.
Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251.
Murphy, P. C., & Sillito, A. M. (1987). Corticofugal feedback influences the generation of length tuning in the visual pathway. *Nature, 329*, 727–729.
Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental sparse and other variants. In M. I. Jordan (Ed.), *Learning in graphical models, 355–368*. Dordrecht: Dordrecht Kulver Academic Press.
Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.
Nordby, H., Hammerborg, D., Roth, W. T., & Hugdahl, K. (1994). ERPs for infrequent omissions and inclusions of stimulus elements. *Psychophysiology, 31*(6), 544–552.
Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *Neuroimage, 50*(1), 302–313.
Rabinovich, M., Huerta, R., & Laurent, G. (2008). Neuroscience: Transient dynamics for neural processing. *Science, 321*(5885), 48–50.
Rao, R. P., & Ballard, D. H. (1998). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience, 2*, 79–87.
Rockland, K. S., & Pandya, D. N. (1979). Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research, 179*, 3–20.
Rohrmeier, M. A., & Koelsch, S. (2012). Predictive information processing in music cognition. *A critical review. Int J Psychophysiol., 83*(2), 164–175.
Rosier, A. M., Arckens, L., Orban, G. A., & Vandesande, F. (1993). Laminar distribution of NMDA receptors in cat and monkey visual cortex visualized by [3H]-MK-801 binding. *Journal of Comparative Neurology, 335*, 369–380.
Sherman, S. M., & Guillery, R. W. (1998). On the actions that one nerve cell can have on another: Distinguishing "drivers" from "modulators". *Proceedings of the National Academy of Sciences USA, 95*, 7121–7126.
Sloboda, J. A. (1991). Music structure and emotional response: Some empirical fmdings. *Psychology of Music, 19*, 110–120.
Steinbeis, N., Koelsch, S., & Sloboda, J. A. (2006). The role of harmonic expectancy violations in musical emotions: evidence from subjective, physiological, and neural responses. *Journal of Cognitive Neuroscience, 18*(8), 1380–1393.
Tsodyks, M. (1999). Attractor neural network models of spatial maps in hippocampus. *Hippocampus, 9*(4), 481–489.
Verleger, R. (1990). P3-evoking wrong notes: unexpected, awaited, or arousing? *International Journal of Neuroscience, 55*, 171–179.
Yabe, H., Tervaniemi, M., Reinikainen, K., & Näätänen, R. (1997). Temporal window of integration revealed by MMN to sound omission. *NeuroReport, 8*(8), 1971–1974.
Zeki, S., & Shipp, S. (1988). The functional logic of cortical connections. *Nature, 335*, 311–331.