Recognizing Sequences of Sequences

Stefan J. Kiebel[1,2]*, Katharina von Kriegstein[1,2], Jean Daunizeau[1], Karl J. Friston[1]

1 Wellcome Trust Centre for Neuroimaging, London, United Kingdom, **2** Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Abstract

The brain's decoding of fast sensory streams is currently impossible to emulate, even approximately, with artificial agents. For example, robust speech recognition is relatively easy for humans but exceptionally difficult for artificial speech recognition systems. In this paper, we propose that recognition can be simplified with an internal model of how sensory input is generated, when formulated in a Bayesian framework. We show that a plausible candidate for an internal or generative model is a hierarchy of 'stable heteroclinic channels'. This model describes continuous dynamics in the environment as a hierarchy of sequences, where slower sequences cause faster sequences. Under this model, online recognition corresponds to the dynamic decoding of causal sequences, giving a representation of the environment with predictive power on several timescales. We illustrate the ensuing decoding or recognition scheme using synthetic sequences of syllables, where syllables are sequences of phonemes and phonemes are sequences of sound-wave modulations. By presenting anomalous stimuli, we find that the resulting recognition dynamics disclose inference at multiple time scales and are reminiscent of neuronal dynamics seen in the real brain.

Citation: Kiebel SJ, von Kriegstein K, Daunizeau J, Friston KJ (2009) Recognizing Sequences of Sequences. PLoS Comput Biol 5(8): e1000464. doi:10.1371/journal.pcbi.1000464

Editor: Rolf Kötter, Radboud University Nijmegen Medical Centre, Netherlands

Received February 18, 2009; **Accepted** July 10, 2009; **Published** August 14, 2009

Funding: SJK is funded by the Max Planck Society. KvK is funded by a independent junior research group grant of the Max Planck Society. JD is funded by a European Marie-Curie fellowship. KJF is funded by the Wellcome Trust. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: skiebel@fil.ion.ucl.ac.uk

Introduction

Many aspects of our sensory environment can be described as dynamic sequences. For example, in the auditory domain, speech and music are sequences of sound-waves [1,2], where speech can be described as a sequence of phonemes. Similarly, in the visual domain, speaking generates sequences of facial cues with biological motion [3,4]. These auditory and visual sequences have an important characteristic: the transitions between the elements are continuous; i.e., it is often impossible to identify a temporal boundary between two consecutive elements. For example, phonemes (speech sounds) in a syllable are not discrete entities that follow each other like beads on a string but rather show graded transitions to the next phoneme. These transitions make artificial speech recognition notoriously difficult [5]. Similarly, in the visual domain, when we observe someone speaking, it is extremely difficult to determine exactly where the movements related to a phoneme start or finish. These dynamic sequences, with brief transitions periods between elements, are an inherent part of our environment, because sensory input is often generated by the fluent and continuous movements of other people, or indeed oneself.

Dynamic sequences are generated on various time-scales. For example, in speech, formants form phonemes and phonemes form syllables. Sequences, which exist at different time-scales, are often structured hierarchically, where sequence elements on one time-scale constrain the expression of sequences on a finer time-scale; e.g. a syllable comprises a specific sequence of phonemes. This functional hierarchy of time-scales may be reflected in the hierarchical, anatomical organisation of the brain [6]. For example, in avian brains, there is anatomical and functional evidence that birdsong is generated and perceived by a hierarchical system, where low levels represent transient acoustic details and high levels encode song structure at slower time-scales [7,8]. An equivalent temporal hierarchy might also exist in the human brain for representing auditory information, such as speech [1,9–12].

Here we ask the following question: How does the brain recognize the dynamic and ambiguous causes of noisy sensory input? Based on experimental and theoretical evidence [13–18] we assume the brain is a recognition system that uses an internal model of its environment. The structure of this model is critical: On one hand, the form of the model must capture the essential architecture of the process generating sensory data. On the other hand, it must also support robust inference. We propose that a candidate that fulfils both criteria is a model based on a hierarchy of stable heteroclinic channels (SHCs). SHCs have been introduced recently as a model of neuronal dynamics *per se* [19]. Here, we use SHCs as the basis of neuronal recognition, using an established Bayesian scheme for modelling perception [20]. This brings together two recent developments in computational approaches to perception: Namely, winnerless competition in stable heteroclinic channels and the hypothesis that the brain performs Bayesian inference. This is important because it connects a dynamic systems perspective on neuronal dynamics [19,21,22] with the large body of work on the brain as an inference machine [13–18].

To demonstrate this we generate artificial speech input (sequences of syllables) and describe a system that can recognize these syllables, online from incoming sound waves. We show that the resulting recognition dynamics display functional characteristics that are reminiscent of psychophysical and neuronal responses.

Author Summary

Despite tremendous advances in neuroscience, we cannot yet build machines that recognize the world as effortlessly as we do. One reason might be that there are computational approaches to recognition that have not yet been exploited. Here, we demonstrate that the ability to recognize temporal sequences might play an important part. We show that an artificial decoding device can extract natural speech sounds from sound waves if speech is generated as dynamic and transient sequences of sequences. In principle, this means that artificial recognition can be implemented robustly and online using dynamic systems theory and Bayesian inference.

Model

In this section, we describe an online recognition scheme for continuous sequences with hierarchical structure. This scheme rests on the concept of stable heteroclinic channels (SHCs) [23], which are combined with an online Bayesian inversion scheme [20]. We now describe these elements and how they are brought together. Note that all variables and their meaning are also listed in Table 1 and 2.

Stable heteroclinic channels (SHCs)

SHCs are attractors formed by artificial neuronal networks, which prescribe sequences of transient dynamics [22–25]. The key aspect of these dynamical systems is that their equations of motion describe a manifold with a series of saddle points. At each saddle point, trajectories are attracted from nearly all directions but are expelled in the direction of another saddle point. If the saddle points are linked up to form a chain, the neuronal state follows a trajectory that passes through all these points, thereby forming a sequence. These sequences are exhibited robustly, even in the presence of high levels of noise. In addition, the dynamics of the SHCs are itinerant due to dynamical instability in the equations of motion and noise on the states. This noise also induces a variation in the exact times that sequence elements are visited. This can be exploited during recognition, where the SHC places prior constraints on the sequence that elements (repelling fixed-points) are visited but does not constrain the exact timing of these visits.

Table 1. Variables used for hierarchies of stable heteroclinic channels (SHCs).

f,g	Nonlinear evolution and observation function
κ	Scalar rate constant
x,v	Hidden and causal state vectors
G_0, β, λ	Scalar control parameters:
	$G_0 = 50$ $\beta = 0.5$ $\lambda = 0.1$
ρ	Inhibitory connectivity matrix
S	Sigmoid function
w,z	state and observation noise vectors
R_k	kth template connectivity matrix

This table lists all variables and their meaning for Eqs. 1 to 3. The additional superscript (j) in Eqs. 2 and 3 denotes the level of the SHC, where level $j=1$ is the lowest.
doi:10.1371/journal.pcbi.1000464.t001

Table 2. Variables used in Bayesian recognition scheme.

y	Sensory input vector
u	Concatenated hidden and causal state vectors $u=\{x,v\}$
m	A model, which specifies the structure of likelihood and priors
$q(u)$	Recognition density used by recognition system to approximate the true but unknown generative density $p(u\|y,m)$
F,U,S	Free energy, energy, and entropy (scalars)
λ	Sufficient statistics vector $\lambda=\{\mu,\Sigma\}$ of normal recognition density q
ε	Prediction error vector (causal states)
e	Prediction error vector (hidden states)

This table lists all variables used in Eqs. 4 to 8. Note that all variables except for m are functions of time.
doi:10.1371/journal.pcbi.1000464.t002

The combination of these two features, robustness of sequence order but flexibility in sequence timing, makes the SHC a good candidate for the neuronal encoding of trajectories [19,26]. Rabinovich et al. have used SHCs to explain how spatiotemporal neuronal dynamics observed in odour perception, or motor control of a marine mollusc, can be expressed in terms of a dynamic system [22,27].

Varona et al. used Lotka-Volterra-type dynamics to model a network of six neurons in a marine mollusc [27]: With particular lateral inhibition between pairs of neurons and input to each neuron, the network displayed sequences of activity. Following a specific order, each neuron became active for a short time and became inactive again, while the next neuron became active, and so on. Stable heteroclinic channels rest on a particular form of attractor manifold that supports itinerant dynamics. This itinerancy can result from deterministic chaos in the absence of noise, which implies the presence of heteroclinic cycles. When noise is added, itinerancy can be assured, even if the original system has stable fixed-points. However, our motivation for considering stochastic differential equations is to construct a probabilistic model, where assumptions about the distribution of noise provide a formal generative model of sensory dynamics.

As reviewed in [22], Lotka-Volterra dynamics can be derived from simple neural mass models of mean membrane potential and mean firing rate [21]. Here, we use a different neural mass model, where the state-vector x can take positive or negative values:

$$\begin{aligned} \dot{x} &= \kappa(-\lambda x - \rho S(x)) + w \\ y &= S(x) + z \\ S(x) &= \frac{G_0}{1+\exp(-\beta x)} \end{aligned} \tag{1}$$

where the motion of a hidden-state vector (e.g., mean membrane potentials) x is a nonlinear function of itself with scalar parameters G_0, β, λ and a connectivity matrix ρ. The hidden state-vector enters a nonlinear function S to generate outcomes (e.g., neuronal firing rates) y. Each element ρ_{ij} determines the strength of lateral inhibition from state j to i. Both the state and observation equations above include additive normally distributed noise vectors w and z. When choosing specific parameter values (see below), the states display stereotyped sequences of activity [28]. Rabinovich et al. [19] termed these dynamics 'stable heteroclinic channels' (SHCs). If the channel forms a ring, once a state is attracted to a saddle point, it will remain in the SHC.

SHCs represent a form of itinerant dynamics [26,29,30] and may represent a substrate for neuronal computations [31]. Remarkably, the formation of SHCs seems to depend largely on the lateral inhibition matrix ρ and not on the type of neuronal model; see Ivanchenko et al. [32] for an example using a complex two-compartment spiking neuron model.

In this paper, we propose to use SHCs not as a model for neuronal dynamics *per se* but as a generative model of how sensory input is generated. This means that we interpret x as hidden states in the environment, which generate sensory input y. The neuronal response to sampling sensory input y are described by recognition dynamics, which decode or deconvolve the causes x from that input. These recognition dynamics are described below. This re-interpretation of Eq. 1 is easy to motivate: sensory input is usually generated by our own body and other organisms. This means input is often generated by neuronal dynamics of the sort described in Eq. 1.

Hierarchies of stable heteroclinic channels

A SHC can generate repetitive, stereotyped sequences. For example, in a system with four saddle points, an SHC forces trajectories through the saddle points in a sequence, e.g. '1-2-3-4-1-2-3-4-1…'. In contrast, a SHC cannot generate '1-2-3-4-3-4-2-1…', because the sequence is not repetitive. However, to model sensory input, for example speech, one must be able to recombine basic sequence-elements like phonemes in ever-changing sequences. One solution would be to represent each possible sequence of phonemes (e.g. each syllable) with a specific SHC. A more plausible and parsimonious solution is to construct a hierarchy of SHCs, which can encode sequences generated by SHCs whose attractor topology (e.g. the channels linking the saddle points) is changed by a supraordinate SHC. This can be achieved by making the connectivity matrix ρ at a subordinate level a function of the output states of the supra-ordinate level. This enables the hierarchy to generate sequences of sequences to any hierarchical depth required.

Following a recent account of how macroscopic cortical anatomy might relate to time-scales in our environment [6], we can construct a hierarchy by setting the rate constant $\kappa^{(j)}$ of the j-th level to a rate that is slower than its subordinate level, $\kappa^{(j-1)}$. As a result, the states of subordinate levels change faster than the states of the level above. This means the control parameters $\rho^{(j)}$ at any level change more slowly than its states, $v^{(j)}$; because the slow change in the attractor manifold is controlled by the supraordinate states:

$$\begin{aligned}\dot{x}^{(j)} &= f^{(j)} + w^{(j)}\\ v^{(j)} &= g^{(j)} + z^{(j)}\\ f^{(j)} &= \kappa^{(j)}\left(-\lambda x^{(j)} - \rho^{(j)}\left(v^{(j+1)}\right)S\left(x^{(j)}\right)\right)\\ g^{(j)} &= S\left(x^{(j)}\right)\end{aligned} \tag{2}$$

where the superscript indexes level j (level 1 being the lowest level), $x^{(j)}$ are 'hidden states', and $v^{(j)}$ are outputs to the subordinate level, which we will call 'causal states'. As before, at the first level, $y = v^{(1)}$ is the sensory stream. In this paper, we consider hierarchies with relative time-scales $\kappa^{(j)}/\kappa^{(j+1)}$ of around four. This means that the time spent in the vicinity of a saddle point at a supraordinate level is long enough for the subordinate level to go through several saddle points. As before, all levels are subject to noise on the motion of the hidden states $w^{(j)}$ and the causal states $z^{(j)}$. At the highest level, the control parameters, $\rho^{(L)}$ are constant over time. At all other levels, the causal states of the supraordinate level, $v^{(j+1)}$, enter the subordinate level by changing the control parameters, the connectivity matrix $\rho^{(j)}$:

$$\rho^{(j)}\left(v^{(j+1)}\right) = \sum_k v_k^{(j+1)} R_k^{(j)} \tag{3}$$

Here, $\rho^{(j)}$ is a linear mixture of 'template' control matrices $R^{(j)}$, weighted by the causal states at level $j+1$. Each of these templates is chosen to generate a SHC. Below, we will show examples of how these templates can be constructed to generate various sequential phenomena. The key point about this construction is that states from the supraordinate level select which template controls the dynamics of the lower level. By induction, the states at each level follow a SHC because the states at the supraordinate level follow a SHC. This means only one state is active at any time and only one template is selected for the lower level. An exception to this is the transition from one state to another, which leads to a transient superposition of two SHC-inducing templates (see below). Effectively, the transition transient at a specific level gives rise to brief spells of non-SHC dynamics at the subordinate levels (see results). These transition periods are characterized by dissipative dynamics, due to the largely inhibitory connectivity matrices, inhibition controlled by parameter λ (Eq. 2) and the saturating nonlinearity S.

In summary, a hierarchy of SHCs generates the sensory stream $y = v^{(1)}$ at the lowest (fastest) level, which forms a sequence of sequences expressed in terms of first-level states. In these models, the lower level follows a SHC, i.e. the states follow an itinerant trajectory through a sequence of saddle points. This SHC will change whenever the supraordinate level, which follows itself a SHC, moves from one saddle point to another. Effectively, we have constructed a system that can generate a stable pattern of transients like an oscillator; however, as shown below, the pattern can have deep or hierarchical structure. Next, we describe how the causes $v^{(j)}$ can be recognized or deconvolved from sensory input y.

Bayesian recognition using SHC hierarchies and the free-energy principle

We have described how SHCs can, in principle, generate sequences of sequences that, we assume, are observed by an agent as its input y. To recognise the causes of the sensory stream the agent must infer the hidden states online, i.e. the system does not look into the future but recognizes the current states x and v of the environment, at all levels of the hierarchy, by the fusion of current sensory input and internal dynamics elicited by past input. An online recognition scheme can be derived from the 'free-energy principle', which states that an agent will minimize its surprise about its sensory input, under a model it entertains about the environment; or, equivalently maximise the evidence for that model [18]. This requires the agent to have a dynamic model, which relates environmental states to sensory input. In this context, recognition is the Bayesian inversion of a generative model. This inversion corresponds to mapping sensory input to the posterior or conditional distribution of hidden states. In general, Bayesian accounts of perception rest on a generative model. Given such a model, one can use the ensuing recognition schemes in artificial perception and furthermore compare simulated recognition dynamics (in response to sensory input), with evoked responses in the brain. The generative model in this paper is dynamical and based on the nonlinear equations 1 and 2. More precisely, these

stochastic differential equations play the role of empirical priors on the dynamics of hidden states causing sensory data.

In the following, we review briefly, the Bayesian model inversion described in [20] for stochastic, hierarchical systems and apply it, in the next section, to hierarchical SHCs.

Given some sensory data vector y, the general inference problem is to compute the model evidence or marginal likelihood of y, given a model m:

$$p(y|m) = \int p(y,u|m)du \tag{4}$$

where the generative model $p(y,u|m) = p(y|u,m)p(u|m)$ is defined in terms of a likelihood $p(y|u,m)$ and prior $p(u|m)$ on hidden states. In Equation 4, the state vector $u = \{x,v\}$ subsumes the hidden and causal states at all levels of a hierarchy (Eq. 2). The model evidence can be estimated by converting this difficult integration problem (Eq. 4) into an easier optimization problem by optimising a free-energy bound on the log-evidence [33]. This bound is constructed using Jensen's inequality and is a function of an arbitrary *recognition* density, $q(u)$:

$$\begin{aligned} F(q,y) &= -\ln p(y|m) + D = U - S \\ D &= \int q(u)\ln\frac{q(u)}{p(u|y,m)}du \geq 0 \end{aligned} \tag{5}$$

The free-energy comprises an energy term $U = -\langle \ln p(y|u) + \ln p(u)\rangle_q$ and an entropy term $S = -\langle \ln q(u)\rangle_q$ and is defined uniquely, given a generative model m. The free-energy is an upper bound on the surprise or negative log-evidence, because the Kullback-Leibler divergence D, between the recognition and conditional density, is always positive. Minimising the free-energy minimises the divergence, rendering the recognition density $q(u)$ an approximate conditional density. When using this approach, one usually employs a parameterized fixed-form recognition density, $q(u|\lambda)$ [20]. Inference corresponds to optimising the free-energy with respect to the sufficient statistics, λ of the recognition density:

$$\begin{aligned} \lambda^* &= \arg\min_{\lambda} F(\lambda,y) \\ q(u|\lambda^*) &\approx p(u|y,m) \end{aligned} \tag{6}$$

The optimal statistics λ^* are sufficient to describe the approximate posterior density; i.e. the agent's belief about (or representation of) the trajectory of the hidden and causal states. We refer the interested reader to Friston et al. [34] for technical details about this variational Bayesian treatment of dynamical systems. Intuitively, this scheme can be thought of as augmented gradient descent on a free-energy bound on the model's log-evidence. Critically, it outperforms conventional Bayesian filtering (e.g., Extended Kalman filtering) and eschews the computation of probability transition matrices. This means it can be implemented in a simple and neuronally plausible fashion [20].

In short, this recognition scheme operates online and recognizes current states of the environment by combining current sensory input with internal recognition dynamics, elicited by past input.

A recognition system that minimizes its free-energy efficiently will come to represent the environmental dynamics in terms of the sufficient statistics of recognition density; e.g. the conditional expectations and variances of $q(u|\lambda) = N(\mu,\Sigma) : \lambda = \{\mu,\Sigma\}$. We assume that the conditional moments are encoded by neuronal activity; i.e., Equation 6 prescribes neuronal recognition dynamics. These dynamics implement Bayesian inversion of the generative model, under the approximations entailed by the form of the recognition density. Neuronally, Equation 6 can be implemented using a message passing scheme, which, in the context of hierarchical models, involves passing prediction errors up and passing predictions down, from one level to the next. These prediction errors are the difference between the causal states (Equation 2);

$$\varepsilon^{(j)} = v^{(j)} - g^{(j)} \tag{7}$$

at any level j, and their prediction from the level above, evaluated at the conditional expectations [18,35]. In addition, there are prediction errors that mediate dynamical priors on the motion of hidden states within each level (Equation 2);

$$e^{(j)} = \dot{x}^{(j)} - f^{(j)} \tag{8}$$

This means that neuronal populations encode two types of dynamics: the conditional expectations of states of the world and the prediction errors. The dynamics of the first are given by Equation 6, which can be formulated as a function of prediction error. These dynamics effectively suppress or explain away prediction error; see [34] for details.

This inversion scheme is a generic recognition process that receives dynamic sensory input and can, given an appropriate generative model, rapidly identify and track environmental states that are generating current input. More precisely, the recognition dynamics resemble the environmental (hidden) states they track (to which they are indirectly coupled), but differ from the latter because they are driven by a gradient descent on free-energy; Eq. 6 (i.e. minimize prediction errors: Eqs. 7 and 8). This is important, because we want to use SHCs as a generative model, not as a model of neuronal encoding *per se*. This means that the neuronal dynamics will only recapitulate the dynamics entitled by SHCs in the environment, if the recognition scheme can suppress prediction errors efficiently in the face of sensory noise and potential beliefs about the world.

We are now in a position to formulate hierarchies of SHCs as generative models, use them to generate sensory input and simulate recognition of the causal states generating that input. In terms of low-level speech processing, this means that any given phoneme will predict the next phoneme. At the same time, as phonemes are recognized, there is also a prediction about which syllable is the most likely context for generating these phonemes. This prediction arises due to the learnt regularities in speech. In turn, the most likely syllable predicts the next phoneme. This means that speech recognition can be described as a dynamic process, on multiple time-scales, with recurrently evolving representations and predictions, all driven by the sensory input.

A model of speech recognition

In the auditory system, higher cortical levels appear to represent features that are expressed at slower temporal scales [36]. Wang et al. [37] present evidence from single-neuron recordings that there is a 'slowing down' of representational trajectories from human auditory sensory thalamus (a 'relay' to the primary auditory cortex), the medial geniculate body (MGB) to primary auditory cortex (AI). In humans, it has been found that the sensory thalamus responds preferentially to faster temporal modulations of sensory signals, whereas primary cortex prefers slower modulations [10]. These findings indicate that neuronal populations, at lower levels of the auditory system (e.g. MGB), represent faster environmental

trajectories than higher levels (e.g., A1). Specifically, the,MGB responds preferentially to temporal modulations of ~20 Hz (~50 ms), whereas AI prefers modulations at ~6 Hz (~150 ms) [10]. Such a temporal hierarchy would be optimal for speech recognition, in which information over longer time-scales provides predictions for processing at shorter time scales. In accord with this conjecture, optimal encoding of fast (rapidly modulated) dynamics by top-down predictions has been found to be critical for communication [1,12,38].

We model this 'slowing down' with a hierarchical generative model based on SHCs. This model generates sequences of syllables, where each syllable is a sequence of phonemes. Phonemes are the smallest speech sounds that distinguishes meaning and a syllable is a unit of organization for a sequence of phonemes. Each phoneme prescribes a sequence of sound-wave modulations which correspond to sensory data. We generated data in this fashion and simulated online recognition (see Figure 1). By recognizing speech-like phoneme-sequences, we provide a proof-of-principle that a hierarchical system can use sensory streams to infer sequences. This not only models the slowing down of representations in the auditory system [10,12,37,38], but may point to computational approaches to speech recognition. In summary, the recognition dynamics following Equation 6 are coupled to a generative model based on SHCs via sensory input. The systems generating and recognising states in Fig. 1 are both dynamic systems, where a non-autonomous recognition system is coupled to an autonomous system generating speech.

All our simulations used hierarchies with two levels (Figure 2). The first (phonemic) level produces a sequence of phonemes, and the second (syllabic) level encodes sequences of syllables. We used Equation 2 to produce phoneme sequences, where the generating parameters are listed in Table 3. The template matrices $R^{(j)}$ (Equation 3) were produced in the following way: We first specified the sequence each template should induce; e.g., sequence 1-2-3 for three neuronal populations. We then set elements on the main diagonal to 1, the elements (2,1), (3,2), (1,3) to value 0.5, and all other elements to 5 [28]. More generally for sequence $s_1, \ldots, s_N$

$$R_{il} = \begin{cases} 1 & i = l \\ .5 & i = s_1, l = s_N \\ .5 & i = s_{n+1}, l = s_n \\ 5 & otherwise \end{cases} \tag{9}$$

Note that SHC hierarchies can be used to create a variety of

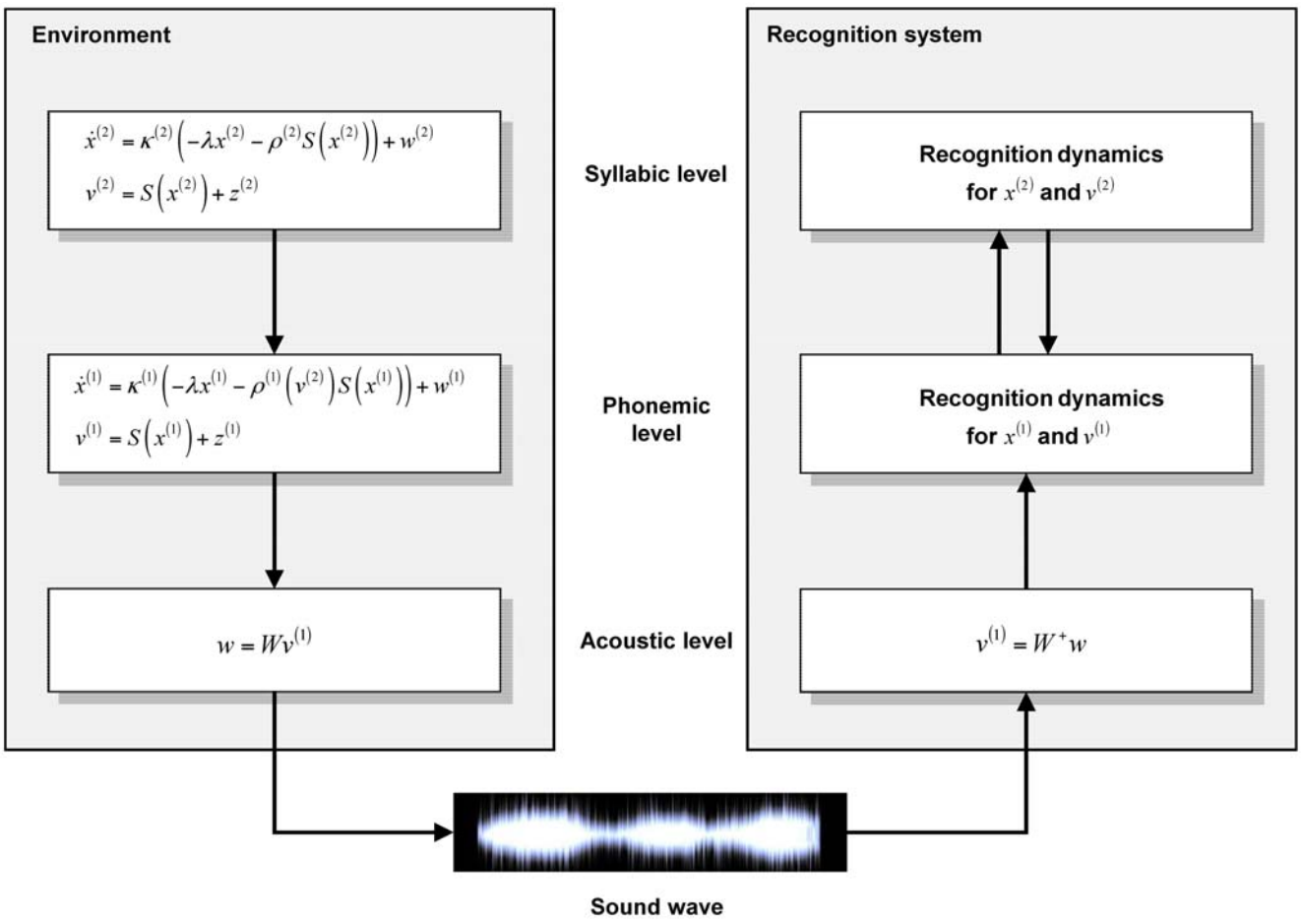

Figure 1. Schematic of the generative model and recognition system. This schematic shows the equations which define both the generation of stimuli (left, see Equation 2) and the recognition scheme based on a generative model. There are three levels; the phonemic and syllabic levels employ stable heteroclinic channels, while the acoustic level is implemented by a linear transform. W corresponds to sound file extracts and w is the resulting sound wave. This sound wave is input to the recognition system, with a linear (forward) projection using the pseudo-inverse W^{+}. The recognition of the phonemic and syllabic level uses bottom-up and top-down message passing between the phonemic and syllabic level, following Equation 6.

doi:10.1371/journal.pcbi.1000464.g001

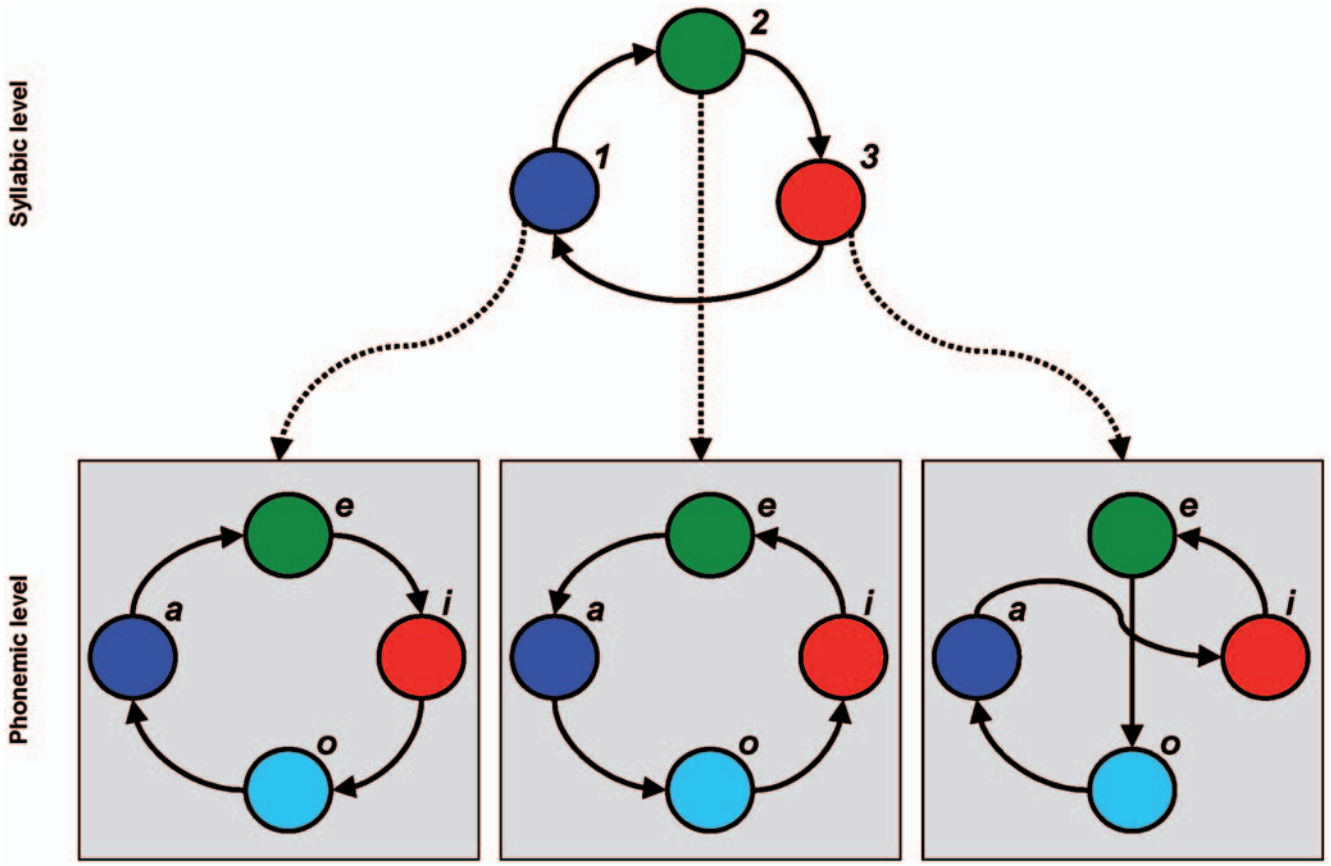

Figure 2. Two-level model to generate phoneme sequences. Schematic illustration of the phoneme sequence generation process. At the syllabic level, one of three syllables is active and induces a specific lateral connectivity structure at the phonemic level. The transition speed at the phonemic level is four times faster than at the syllabic level. The resulting phoneme and syllable dynamics of the model are shown in Fig. 3a.
doi:10.1371/journal.pcbi.1000464.g002

different behaviours, using different connectivity matrices. Here we explore only a subset of possible sequential dynamics.

When generating sensory data y, we added noise $w^{(j)}$ and $z^{(j)}$ to both the hidden and causal states. At the first and second levels, this was normally distributed zero-mean noise with log-precisions of ten and sixteen, respectively. These noise levels were chosen to introduce noisy dynamics but not to the extent that the recognition became difficult to visualise. We repeated all the simulations reported below with higher noise levels and found that the findings remained qualitatively the same (results not shown). Synthetic stimuli were generated by taking a linear mixture of sound waves extracted from sound files, in which a single speaker pronounced each of four vowel-phonemes: [a], [e], [i], [o]. These extracts W were sampled at 22050 Hz and about 14 ms long. The mixture was weighted by the causal states of the phonemic level; $w = Wv^{(1)}$. This resulted in a concatenated sound wave file w. When this sound file is played, one perceives a sequence of vowels with smooth, overlapping transitions (audio file S1). These transitions are driven by the SHCs guiding the expression of the phonemes and syllables at both levels of the generative hierarchy.

For computational simplicity, we circumvented a detailed generative model of the acoustic level. For simulated recognition, the acoustic input (the sound wave) was transformed to phonemic input by inverting the linear mixing described above every seven ms of simulated time (one time bin). This means that our recognition scheme at the acoustic level assumes forward processing only (Fig. 1). However, in principle, given an appropriate generative model [39,40], one could invert a full acoustic model, using forward and backward message passing between the acoustic and phonemic levels.

Results

In this section, we illustrate that the recognition scheme described above can reliably decode syllabic and phonemic structure from sensory input online, if it has the correct generative model. We will also describe how recognition fails when the generative model does not have a form that provides veridical predictions of the sensorium, e.g., when agents are not conspecific or we hear a foreign language. These simulations relate to empirical studies of brain responses evoked by unpredicted linguistic stimuli. We conclude with a more subtle violation that we deal with in everyday audition; namely the recognition of speech presented at different speeds.

Recognising a sequence of sequences

To create synthetic stimuli we generated syllable sequences consisting of four phonemes or states; [a], [e], [i], and [o], over 11.25 seconds (800 time points), using a two-level SHC model (Fig. 2). To simulate word-like stimuli, we imposed silence at the beginning and the end by windowing the phoneme sequence (Fig. 3A, top left). At the syllabic level, we used three syllables or states to form the second-level sequence $(1\text{–}2\text{–}3)^{(2)}$; where the numbers denote the sequence and the superscript indicates the

Table 3. Default parameters used for simulations with Equations 2 and 3.

λ	0.3
G_0	50
β	0.5
$\kappa^{(1)}$	1/8
$\kappa^{(2)}$	1/32

doi:10.1371/journal.pcbi.1000464.t003

sequence level. The three causal states $v^{(2)}$ of the syllabic level entered the phonemic level as control parameters to induce their template matrices as in Equation 3. This means that each of the three syllable states at the second level causes a phoneme sequence at the first: $(a-e-i-o)^{(1)}$, $(o-i-e-a)^{(1)}$, and $(a-i-e-o)^{(1)}$, see Fig. 2 and listen to the audio file S1. In Fig. 3A we show the causal and hidden states, at both levels, generated by this model. The remaining parameters, for both levels, are listed in Table 3. Note that the rate constant of the syllabic level is four times slower than at the phonemic level. As expected, the phoneme sequence at the first level changes as a function of the active syllable at the second level. The transients caused by transitions between syllables manifest at the first level as temporary changes in the amplitude or duration of the active phoneme.

We then simulated recognition of these sequences. Fig. 3B shows that our recognition model successfully tracks the true states at both levels. Note the recognition dynamics rapidly 'lock onto' the causal states from the onset of the first phoneme of the first syllable (time point 50). Interestingly, the system did not recognize the first syllable (true: syllable 3 (red line), recognized: syllable 2 (green line) between time points 50 to 80 (see red arrow in Fig. 3B), but corrected itself fairly quickly, when the sensory stream indicated a new phoneme that could only be explained by the third syllable. This initial transient at the syllabic level shows that recognition dynamics can show small but revealing deviations from the true state dynamics. In principle, these deviations could be used to test whether the real auditory system uses a recognition algorithm similar to the one proposed; in particular, the simulated

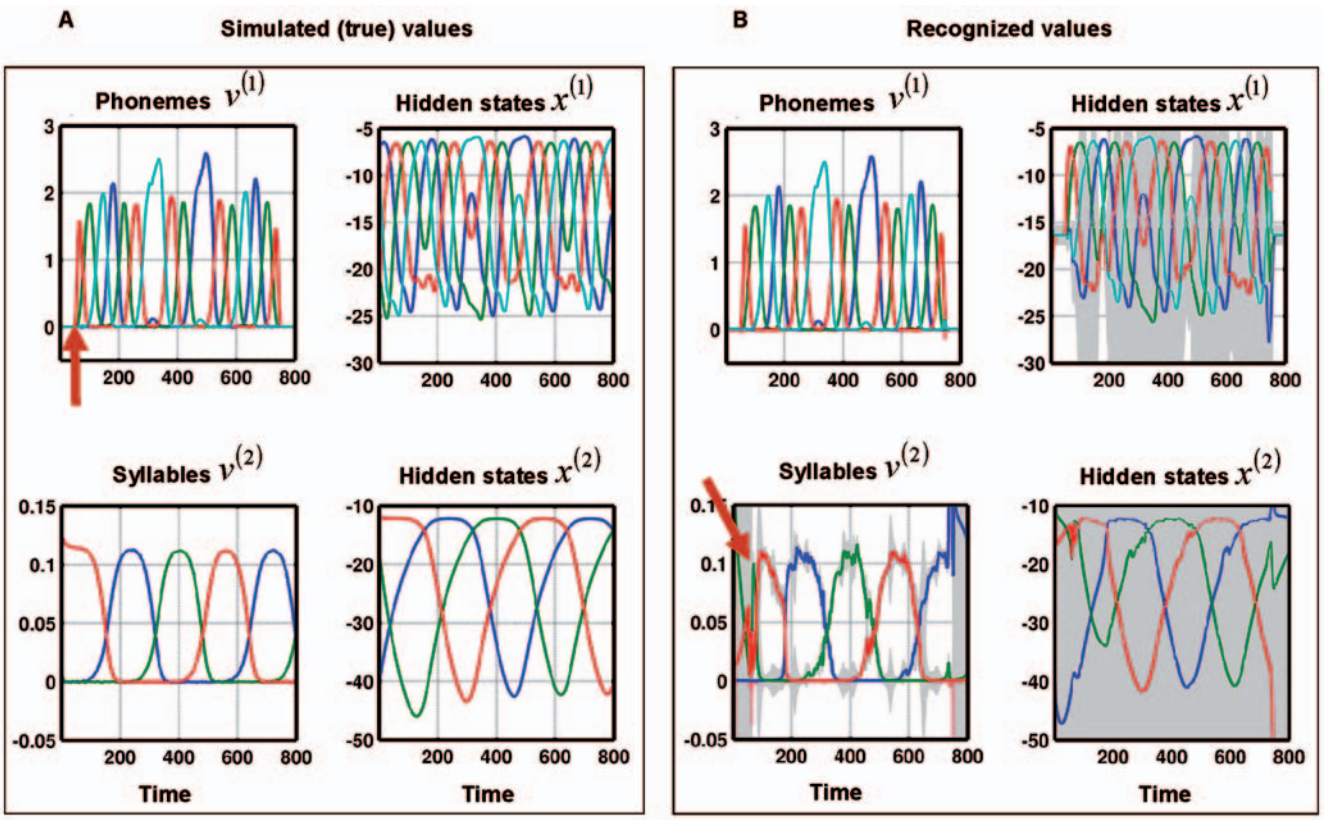

Figure 3. Recognition of a sequence of sequences. (A): Dynamics of generated causal and hidden states at the phonemic and syllabic level, using Equation 2. At the syllabic level, there are three different syllables (1: blue, 2: green, 3: red), following the syllable sequence 1→2→3. The slowly changing state Syllable 1 causes the faster-moving phoneme sequence a→e→i→o (blue→green→red→cyan), syllable 2: o→i→e→a (cyan→red→green→blue), and syllable 3: a→i→e→o (blue→red→green→cyan). See Fig 2 for a schematic description of these sequences. At the beginning and end of the time-series $v^{(1)}$ (top-left plot), we introduced silence by applying a windowing function to zero time points 0 to 50 and 750 to 800. The red arrow indicates the end of the initial silent period. The phonemic states $v^{(1)}$ cause sound waves, resolved at 22050 Hz (see Fig. 1). These sound waves are the input to the recognition system. (B): The recognition dynamics after inverting the sound wave. At the phonemic level, the states follow the true states closely. At the syllabic level, the recognized causal state dynamics $v^{(2)}$ are rougher than the true states but track the true syllable sequence veridically. The high-amplitude transients of $v^{(2)}$ at the beginning and end of the time-series are due to the silent periods, where the syllabic recognition states $v^{(2)}$ experience high uncertainty (plotted in grey: confidence intervals of 95% around the mean). Note that the hidden states, at both levels, experience high uncertainty whenever a phoneme or syllable is inactive. The red arrow indicates an initial but rapidly corrected mis-recognition of the causing syllable.
doi:10.1371/journal.pcbi.1000464.g003

recognition dynamics could be used to explain empirical neurophysiological responses.

Sensitivity to sequence violations

What happens if the stimuli deviate from learned expectations (e.g. violation of phonotactic rules)? In other words, what happens if we presented known phonemes that form unknown syllables? This question is interesting for two reasons. First, our artificial recognition scheme should do what we expect real brains to do when listening to a foreign language: they should be able to recognize the phonemes but should not derive high-order 'meaning' from them; i.e. should not recognize any syllable. Secondly, there are well-characterised brain responses to phonotactic violations, e.g. [41–43]. These are usually event-related responses that contain specific waveform components late in peristimulus time, such as the N400. The N400 is an event-related potential (ERP) component typically elicited by unexpected linguistic stimuli. It is characterized as a negative deflection (topologically distributed over central-parietal sites on the scalp), peaking approximately 400 ms after the presentation of an unexpected stimulus.

To model phonotactic violations, we generated data with the two-level model presented above. However, we used syllables, i.e. sequences of phonemes, that the recognition scheme was not informed about and consequently could not recognise (it has three syllables in its repertoire: $(a-i-o-e)^{(1)}$, $(a-o-e-i)^{(1)}$, and $(a-e-o-i)^{(1)}$). Thus the recognition scheme knows all four phonemes but is unable to predict the sequences heard. Fig. 4A shows that the recognition system cannot track the syllables; the recognized syllables are very different from the true syllable dynamics. At the phonemic level, the prediction error $e^{(1)}$ deviates from zero whenever a new (unexpected) phoneme is encountered (Fig. 4B). The prediction error at the syllabic level is sometimes spike-like and can reach high amplitudes, relative to the typical amplitudes of the true states (see Fig. 4A and B). This means that the prediction error signals violation of phonotactic rules. In Fig. 4C, we zoom in onto time points 440 to 470 to show how the prediction error evolves when evidence of a phonotactic violation emerges: At the phoneme level, prediction error builds up because an unexpected phoneme appears. After time point 450, the prediction error $e^{(1)}$ grows quickly, up to the point that the system resolves the prediction error. This is done by 'switching' to a new syllable, which can explain the transition to the emerging phoneme. The switching creates a large amplitude prediction error $e^{(2)}$ at time point 460. In other words, in face of emerging evidence that its current representation of syllables and phonemes cannot explain sensory input, the system switches rapidly to a new syllable representation, giving rise to a new prediction error. It

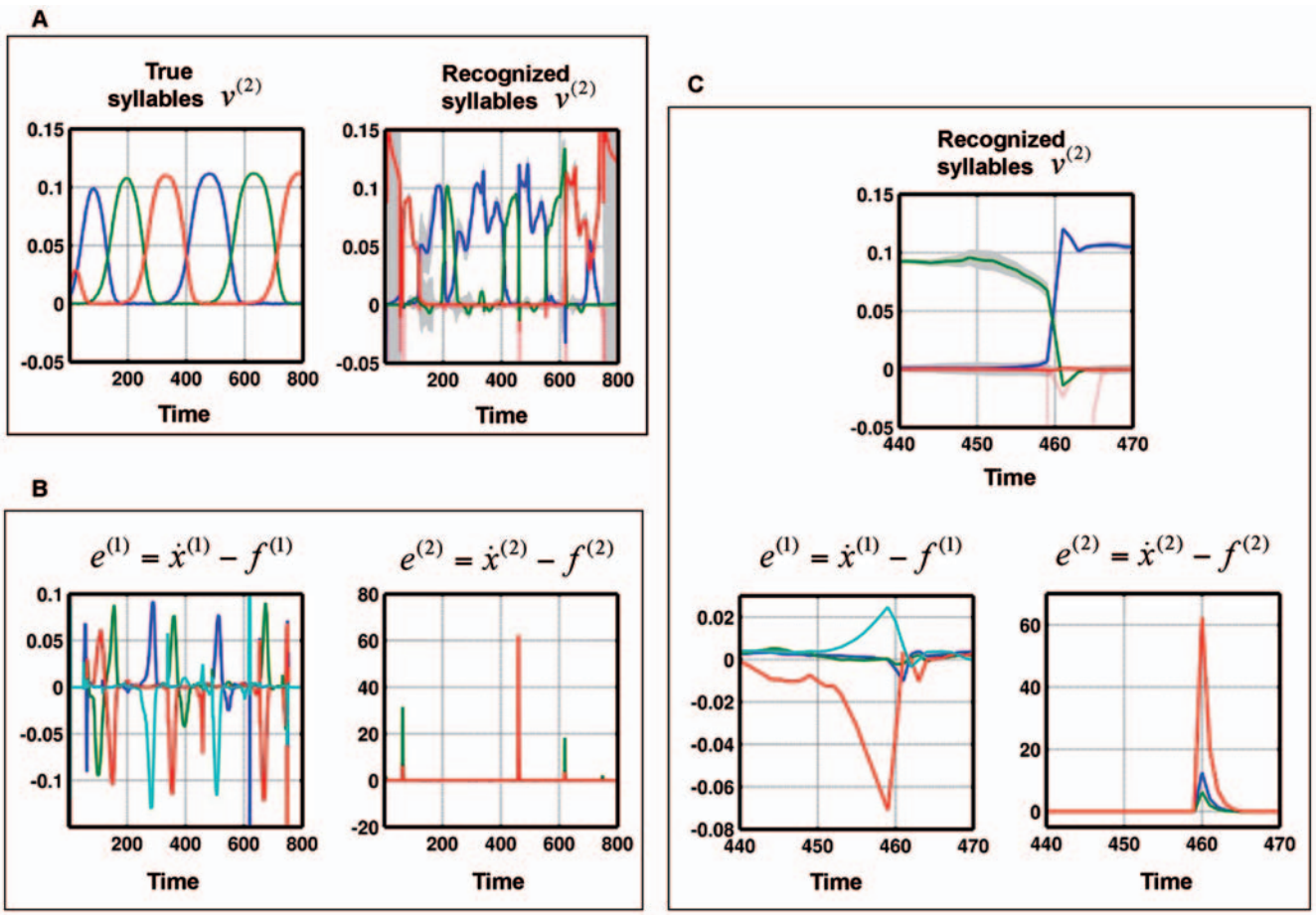

Figure 4. Recognition of sequences with phonotactic violation. (A): True and recognized syllable dynamics of a two-level model when the syllables are unknown to the recognition system. Left: True dynamics of $v^{(2)}$, Right: Recognition dynamics for $v^{(2)}$. (B): Left: Prediction error $e^{(1)}$ at phonemic level. Right: Prediction error $e^{(2)}$ at syllabic level. (C): Zoom of dynamics shown in A and B from time points 440 to 470. See text for description of these dynamics.
doi:10.1371/journal.pcbi.1000464.g004

may be that these prediction errors are related to electrophysiological responses to violations of phonotactic rules, [44,45]. This is because the largest contributors to non-invasive electromagnetic signals are thought to be superficial pyramidal cells. In biological implementations of the recognition scheme used here [20], these cells encode prediction error.

In summary, these simulations show that a recognition system cannot represent trajectories or sequences that are not part of its generative model. In these circumstances, recognition experiences intermittent high-amplitude prediction errors because the internal predictions do not match the sensory input. There is a clear formal analogy between the expression of prediction error in these simulations and mismatch or prediction violation responses observed empirically. The literature that examines event-related brain potentials (ERPs) and novelty processing "reveals that the orienting response engendered by deviant or unexpected events consists of a characteristic ERP pattern, comprised sequentially of the mismatch negativity (MMN) and the novelty P3 or P3a" [46].

Robustness to speed of speech

Human speech recognition is robust to the speed of speech [47,48]. How do our brains recognize speech at different rates? There are two possible mechanisms in our model that can deal with 'speaker speed' parameters online. First, one could make the rate constants $\kappa^{(1)}$ and $\kappa^{(2)}$ free parameters and optimise them during inversion. Adjusting to different speaker parameters is probably an essential faculty, because people speak at different speeds [49]. The second mechanism is that the recognition itself might be robust to deviations from the expected rate of phonemic transitions; i.e., even though the recognition uses the rate parameters appropriate for much slower speech, it still can recognize fast speech. This might explain why human listeners can understand speech at rates that they have never experienced previously [47]. In the following, we show that our scheme has this robustness.

To simulate speed differences we used the same two-level model as in the simulations above with $\kappa^{(1)} = 1/8$ for the generation of phonemes, but with $\kappa^{(1)} = 1/12$ for recognition so that the stimulus stream was 50% faster than expected. As can be seen in Fig. 5A, the recognition can successfully track the syllables. This was because the second level supported the adaption to the fast sensory input by changing its recognition dynamics in responses to prediction error (see Fig. 5B: note the amplitude difference in Fig. 5A between the true and recognized $v^{(2)}$). The prediction errors at both levels, $e^{(1)}$ and $e^{(2)}$, are shown in Fig. 5C. In particular, the second-level error $e^{(2)} = \dot{x}^{(2)} - f^{(2)}$ displayed spike-like corrections around second-level transitions. These are small in amplitude compared to both the amplitude of the hidden states and the prediction errors of the previous simulation (Fig. 4B). These results show that the system can track the true syllables veridically, where the prediction error accommodates the effects caused by speed differences. This robustness to variations in the speed of phoneme transitions might be a feature shared with the auditory system [50].

Discussion

We have shown that stable heteroclinic channels (SHCs) can be used as generative models for online recognition. In particular, we have provided proof-of-concept that sensory input generated by these hierarchies can be deconvolved to disclose the hidden states causing that input. This is a non-trivial observation because nonlinear, hierarchical and stochastic dynamical systems are difficult to invert online [51,52]. However, we found that the inversion of models based on SHCs is relatively simple. Furthermore, the implicit recognition scheme appears robust to noise and deviations from true parameters. This suggests that SHCs may be a candidate for neuronal models that contend with the same problem of de-convolving causes from sensory consequences. Moreover, hierarchical SHCs seem, in principle, an appropriate description of natural sequential input, which is usually generated by our own body or other organisms, and can be described as a mixture of transients and discrete events.

The general picture of recognition that emerges is as follows: Sensory input is generated by a hierarchy of dynamic systems in the environment. We couple this dynamic system, via sensory sampling, to our recognition system implementing the inversion dynamics (Fig. 1). The recognition system minimizes a proxy for surprise or model evidence; the negative free-energy (Eq. 6). To do this, the states of the recognition system move on manifolds, defined through the free-energy by the generative model. Here, we use a hierarchy of SHCs as generative model so that the manifold changes continuously at various time-scales. The inferred SHC states never reach a fixed point, but are perpetually following a trajectory through state-space, in the attempt to mirror the generative dynamics of the environment. When sensory input is unexpected (see second simulation, Fig. 4), the system uses the prediction error to change its representation quickly, at all levels, such that it best explains the sensory stream.

In a previous paper [6], we have shown that one can use chaotic attractors (i.e., a hierarchy of Lorenz attractors) to model auditory perception. However, SHCs may provide a more plausible model of sensory dynamics: First, they show structure over extended temporal scales, much like real sensory streams. This may reflect the fact that the processes generating sensory data are themselves (usually) neuronal dynamics showing winnerless competition. Secondly, many chaotic systems like the Lorenz attractor have only few states and cannot be extended to high dimensions in a straightforward fashion. This was no problem in our previous model, where we modelled a series of simple chirps, with varying amplitude and frequency [6]. However, it would be difficult to generate sequences of distinct states that populate a high dimensional state-space; e.g. phonemes in speech. In contrast, stable heteroclinic channels can be formulated easily in high dimensional state spaces.

In this paper, we used a generative model which was formally identical to the process actually generating sensory input. We did this for simplicity; however, any generative model that could predict sensory input would be sufficient. In one sense, there is no true model because it is impossible to disambiguate between models that have different forms but make the same predictions. This is a common issue in ill-posed inverse problems, where there are an infinite number of models that could explain the same data. In this context the best model is usually identified as the most parsimonious. Furthermore, we are not suggesting that all aspects of perception can be framed in terms of the inversion of SHCs; we only consider recognition of those sensory data that are generated by mechanisms that are formally similar to the itinerant but structured dynamics of SHCs.

The proof-of-concept presented above makes the SHC hierarchy a potential candidate for speech recognition models. The recognition dynamics we simulated can outpace the dynamics they are trying to recognise. In all our simulations, after some initial transient, the recognition started tracking the veridical states early in the sequence. For example, the scheme can identify the correct syllable before all of its phonemes have been heard. We only simulated two levels, but this feature of fast recognition on exposure to brief parts of the sequence may hold for many more levels. Such rapid recognition of potentially long sequences is seen

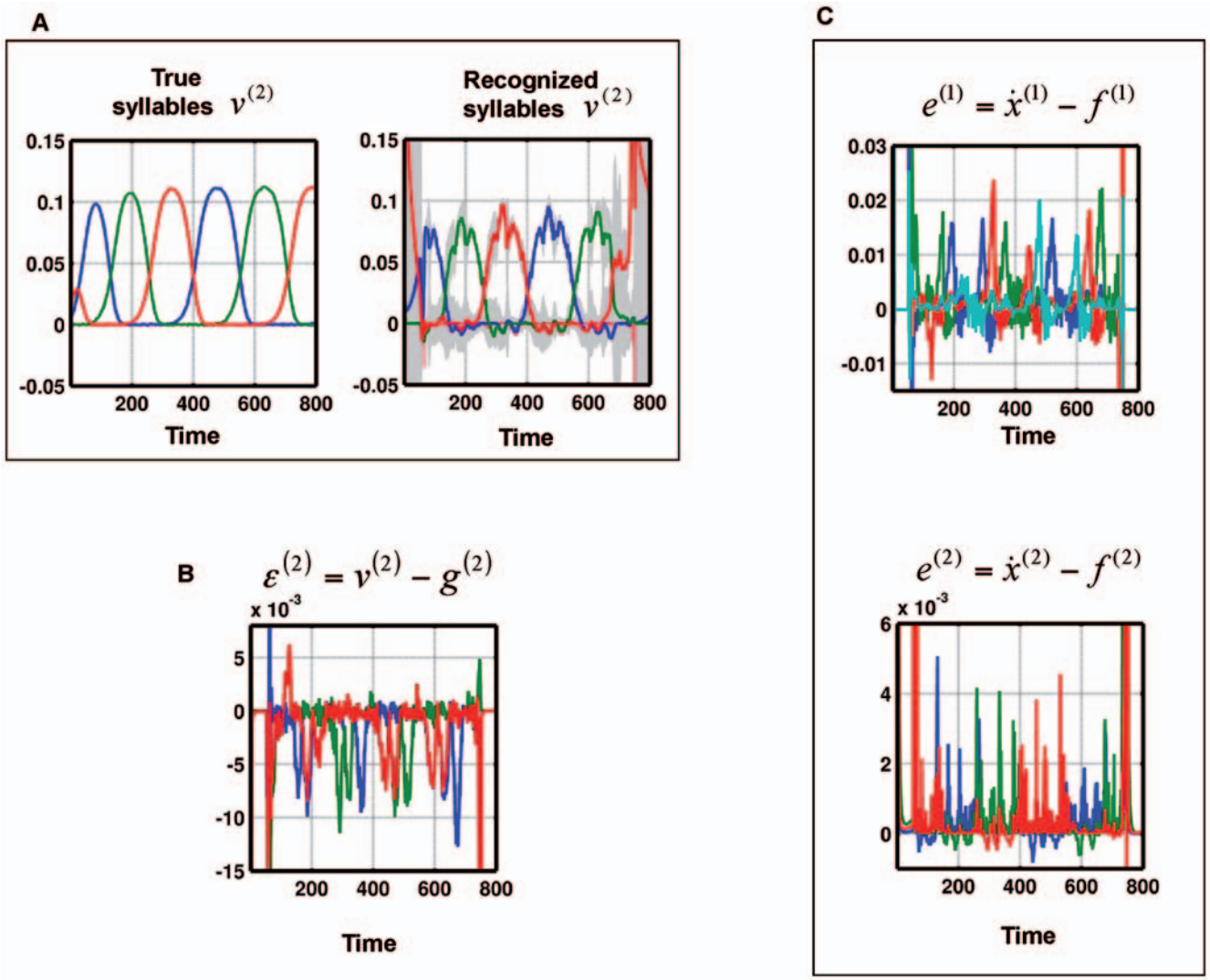

Figure 5. Recognition of unexpectedly fast phoneme sequences. (A): True and recognized syllable dynamics of a two-level model when the phoneme sequence is generated with a rate constant of $\kappa^{(1)} = 1/8$ but recognized with a rate constant of $\kappa^{(1)} = 1/12$, i.e. speech was 50% faster than expected. Left: True dynamics of $v^{(2)}$, Right: Recognition dynamics for $v^{(2)}$. (B): Prediction error $\varepsilon^{(2)}$ at syllabic level. (C) Top: Prediction error $e^{(1)}$ at phonemic level. Bottom: Prediction error $e^{(2)}$ at syllabic level.
doi:10.1371/journal.pcbi.1000464.g005

in real systems; e.g., we can infer that someone is making a cup of tea from observing a particular movement, like getting a teabag out of a kitchen cupboard. The reason why recognition can be fast is that the generative model is nonlinear (through the top-down control of attractor manifolds). With nonlinearities, slow time-scales in hierarchical sequences can be recognized rapidly because they disclose themselves in short unique sequences in the sensory input. Furthermore, we demonstrated another requirement for efficient communication: recognition signals, via prediction error, when unrecognised syllables cannot be decoded with its phonotactic model. This is important, because, an agent can decide online whether its decoding of the message is successful or not. Following the free-energy principle, this would oblige the agent to act on its environment, so that future prediction error is minimized [18]. For example, the prediction error could prompt an action ('repeat, please') and initiate learning of new phonotactic rules.

Another aspect of SHC-based models is that they can recombine sensory primitives like phonemes in a large number of ways. This means that neuronal networks implementing SHC dynamics, based on a few primitives at the first level, can encode a large number of sequences. This feature is critical for encoding words in a language; e.g., every language contains many more words than phonemes [53]. The number of sequences that a SHC system can encode is

$$\sum_{k=3}^{N} \binom{N}{k} (k-1)! \tag{10}$$

where N is the number of elements [22]. This would mean, in theory, that the number of states that can be encoded with a sequence, given a few dozens primitives, is nearly endless. It is unlikely that this full capacity is exploited in communication. Rather, for efficient communication, it might be useful to restrict the number of admissible sequences to make them identifiable early in the sequence.

We did not equip the recognition model with a model of the silent periods at the beginning and end of a word (Fig. 3A). It is interesting to see how recognition resolves this: to approximate silence, the

system held hidden phoneme states very negative by driving the states away from the SHC attractor and tolerating the violation of top-down predictions. However, the tolerance is limited as can be seen by the slightly positive inferred hidden states (Fig. 3B). Such behaviour is beneficial for recognition because the agent, within bounds, can deviate from internal predictions. A built-in error tolerance which is sensitive to the kind of errors it should endure to make recognition robust is important in an uncertain world. Robustness to errors would be impossible with an inversion scheme based on a deterministic model, which assumes that the sensory input follows a deterministic trajectory without any noise on the environmental causes. With such a recognition system, the agent could not deal with (unexpected) silence, because the SHC-based inversion dynamics would attract the state-trajectory without any means of resolving the resulting prediction error between the zero (silent) sensory input and the internal predictions. Recognition schemes based on stochastic systems can deviate adaptively from prior predictions, with a tolerance related to the variance of the stochastic innovations. Optimising this second-order parameter then becomes critical for recognition (see [20]).

Links to neuroscience

There is emerging evidence in several areas of neuroscience that temporal hierarchies play a critical role in brain function [6]. The three areas where this is most evident are auditory processing [12,37,54–56], cognitive control [57–59], and motor control [60]. Our conclusions are based on a generic recognition scheme [20] and are therefore a consequence of our specific generative model, a temporal hierarchy of SHCs. This hierarchy of time-scales agrees well with the temporal anatomy of the hierarchical auditory system, where populations close to the periphery encode the fast acoustics, while higher areas form slower representations [9,10,37,38,61,62]. In particular, our model is consistent with findings that phonological (high) levels have strong expectations about the relevance of acoustic (low) dynamics [38].

Neurobiological treatments of the present framework suppose that superficial pyramidal cell populations encode prediction error; it is these cells that contribute most to evoked responses as observed in magneto/electroencephalography (M/EEG) [63]. There is an analogy between the expression of prediction error in our simulations and mismatch or prediction violation responses observed empirically. In our simulations, prediction error due to a deviation from expectations is resolved by all levels (Fig. 4B). This might be an explanation for prominent responses to prediction violations to be spatially distributed, e.g., the mismatch negativity, the P300, and the N400 all seem to involve various brain sources in temporal and frontal regions [45,46,64–66]. Inference on predictable auditory streams has been studied and modelled in several ways, in an attempt to explain the rapid recognition of words in the context of sentences, e.g., [38,67–70]. Our simulations show how, in principle, these accounts might be implemented in terms of neuronal population dynamics.

Links to computational models

Learning, storing, inferring and executing sequences is a key topic in experimental [71–79], and theoretical neurosciences [80–82]; and robotics [83–86]. An early approach to modelling sequence processing focussed on feed-forward architectures. However, it was realised quickly that these networks could not store long sequences, because new input overwrote the internal representation of past states. The solution was to introduce explicit memory into recurrent networks, in various forms; e.g. as contextual nodes or 'short-term memory' [87,88]. Although framed in different terms, these approaches can be seen as an approximation to temporal hierarchies, where different units encode representations at different time-scales.

A central issue in modelling perception is how sequences are not just recalled but used as predictions for incoming sensory input. This requires the 'dynamic fusion' of bottom-up sensory input and top-down predictions, Several authors e.g., [83,89–92] use recurrent networks to implement this fusion. Exact Bayesian schemes based on discrete hierarchical hidden Markov models, specified as a temporal hierarchy, have been used to implement memory and recognition [93]. Here, we have used the free-energy principle (i.e. variational Bayesian inference on continuous hierarchical dynamical systems) to show how the ensuing recognition process leads naturally to a scheme which can deal with fast sequential inputs.

In conclusion, we have described a scheme for inferring the causes of sensory sequences with hierarchical structure. The key features of this scheme are: (i) the ability to describe natural sensory input as hierarchical and dynamic sequences, (ii) modeling this input using generative models, (iii) using dynamic systems theory to create plausible models, and (iv) online Bayesian inversion of the resulting models. This scheme is theoretically principled but is also accountable to the empirical evidence available from the auditory system; furthermore, the ensuing recognition dynamics are reminiscent of real brain responses.

Supporting Information

Audio S1 Phoneme sequence generated in first simulation - mpg-file containing phoneme sequence sampled at 22050 Hz. The time courses of the four vowels can be seen in Fig. 3A (top left). Found at: doi:10.1371/journal.pcbi.1000464.s001 (0.18 MB MPG)

Author Contributions

Conceived and designed the experiments: SJK. Performed the experiments: SJK. Analyzed the data: SJK. Contributed reagents/materials/analysis tools: SJK JD KJF. Wrote the paper: SJK KvK JD KJF.

References

1. Poeppel D, Idsardi WJ, van WV (2008) Speech perception at the interface of neurobiology and linguistics. PhilosTransRSocLond B BiolSci 363: 1071–1086.
2. Zatorre RJ, Belin P, Penhune VB (2002) Structure and function of auditory cortex: music and speech. Trends Cogn Sci 6: 37–46.
3. Simon D, Craig KD, Miltner WH, Rainville P (2006) Brain responses to dynamic facial expressions of pain. Pain 126: 309–318.
4. Thompson JC, Hardee JE, Panayiotou A, Crewther D, Puce A (2007) Common and distinct brain activation to viewing dynamic sequences of face and hand movements. Neuroimage 37: 966–973.
5. Deng L, Yu D, Acero A (2006) Structured speech modeling. Ieee Transactions on Audio Speech and Language Processing 14: 1492–1504.
6. Kiebel SJ, Daunizeau J, Friston KJ (2008) A hierarchy of time-scales and the brain. PLoS ComputBiol 4: e1000209.
7. Long MA, Fee MS (2008) Using temperature to analyse temporal dynamics in the songbird motor pathway. Nature 456: 189–194.
8. Sen K, Theunissen FE, Doupe AJ (2001) Feature analysis of natural sounds in the songbird auditory forebrain. JNeurophysiol 86: 1445–1458.
9. Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. JNeurosci 23: 3423–3431.
10. Giraud AL, Lorenzi C, Ashburner J, Wable J, Johnsrude I, et al. (2000) Representation of the temporal envelope of sounds in the human brain. JNeurophysiol 84: 1588–1598.
11. Overath T, Kumar S, von Kriegstein K, Griffiths TD (2008) Encoding of spectral correlation over time in auditory cortex. JNeurosci 28: 13268–13273.
12. von Kriegstein K, Patterson RD, Griffiths TD (2008) Task-dependent modulation of medial geniculate body is behaviorally relevant for speech recognition. CurrBiol 18: 1855–1859.

13. Wolpert DM, Ghahramani Z, Jordan MI (1995) An internal model for sensorimotor integration. Science 269: 1880–1882.
14. Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. Annual Review of Psychology 55: 271–304.
15. Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? Trends in Cognitive Sciences 10: 301–308.
16. Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America a-Optics Image Science and Vision 20: 1434–1448.
17. Kording KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. Nature 427: 244–247.
18. Friston K, Kilner J, Harrison L (2006) A free energy principle for the brain. JPhysiol Paris 100: 70–87.
19. Rabinovich M, Huerta R, Laurent G (2008) Neuroscience - Transient dynamics for neural processing. Science 321: 48–50.
20. Friston K (2008) Hierarchical models in the brain. PLoS ComputBiol 4: e1000211.
21. Fukai T, Tanaka S (1997) A simple neural network exhibiting selective activation of neuronal ensembles: from winner-take-all to winners-share-all. Neural Comput 9: 77–97.
22. Rabinovich MI, Varona P, Selverston AI, Abarbanel HDI (2006) Dynamical principles in neuroscience. Reviews of Modern Physics 78: 1213–1265.
23. Rabinovich MI, Huerta R, Varona P, Afraimovich VS (2008) Transient cognitive dynamics, metastability, and decision making. Plos Computational Biology 4.
24. Afraimovich VS, Zhigulin VP, Rabinovich MI (2004) On the origin of reproducible sequential activity in neural circuits. Chaos 14: 1123–1129.
25. Rabinovich M, Volkovskii A, Lecanda P, Huerta R, Abarbanel HDI, et al. (2001) Dynamical encoding by networks of competing neuron groups: Winnerless competition. Physical Review Letters 8706.
26. Friston KJ (1997) Transients, metastability, and neuronal dynamics. Neuroimage 5: 164–171.
27. Varona P, Rabinovich MI, Selverston AI, Arshavsky YI (2002) Winnerless competition between sensory neurons generates chaos: A possible mechanism for molluscan hunting behavior. Chaos 12: 672–677.
28. Afraimovich VS, Rabinovich MI, Varona P (2004) Heteroclinic contours in neural ensembles and the winnerless competition principle. International Journal of Bifurcation and Chaos 14: 1195–1208.
29. Breakspear M, Terry JR, Friston KJ (2003) Modulation of excitatory synaptic coupling facilitates synchronization and complex dynamics in a nonlinear model of neuronal dynamics. Neurocomputing 52-4: 151–158.
30. Durstewitz D, Deco G (2008) Computational significance of transient dynamics in cortical networks. European Journal of Neuroscience 27: 217–227.
31. Buonomano DV, Maass W (2009) State-dependent computations: spatiotemporal processing in cortical networks. NatRevNeurosci 10: 113–125.
32. Ivanchenko MV, Nowotny T, Selverston AI, Rabinovich MI (2008) Pacemaker and network mechanisms of rhythm generation: Cooperation and competition. Journal of Theoretical Biology 253: 452–461.
33. Beal MJ (2003) Variational algorithms for approximate Bayesian inference [PhD]: University of London.
34. Friston KJ, Trujillo-Barreto N, Daunizeau J (2008) DEM: a variational treatment of dynamic systems. Neuroimage 41: 849–885.
35. Friston K (2005) A theory of cortical responses. Philosophical Transactions of the Royal Society B-Biological Sciences 360: 815–836.
36. Creutzfeldt O, Hellweg FC, Schreiner C (1980) Thalamocortical Transformation of Responses to Complex Auditory-Stimuli. Experimental Brain Research 39: 87–104.
37. Wang X, Lu T, Bendor D, Bartlett E (2008) Neural coding of temporal information in auditory thalamus and cortex. Neuroscience 154: 294–303.
38. Nahum M, Nelken I, Ahissar M (2008) Low-level information and high-level perception: The case of speech in noise. Plos Biology 6: 978–991.
39. Holmberg M, Gelbart D, Hemmert W (2007) Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition. Speech Communication 49: 917–932.
40. Sumner CJ, Lopez-Poveda EA, O'Mard LP, Meddis R (2002) A revised model of the inner-hair cell and auditory-nerve complex. Journal of the Acoustical Society of America 111: 2178–2188.
41. Dehaene-Lambertz G, Dupoux E, Gout A (2000) Electrophysiological correlates of phonological processing: A cross-linguistic study. Journal of Cognitive Neuroscience 12: 635–647.
42. Eulitz C, Lahiri A (2004) Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. Journal of Cognitive Neuroscience 16: 577–583.
43. Friedrich M, Friederici AD (2005) Phonotactic knowledge and lexical-semantic processing in one-year-olds: Brain responses to words and nonsense words in picture contexts. Journal of Cognitive Neuroscience 17: 1785–1802.
44. Friederici AD (2002) Towards a neural basis of auditory sentence processing. Trends in Cognitive Sciences 6: 78–84.
45. Lau EF, Phillips C, Poeppel D (2008) A cortical network for semantics: (de)constructing the N400. Nature Reviews Neuroscience 9: 920–933.
46. Friedman D, Cycowicz YM, Gaeta H (2001) The novelty P3: an event-related brain potential (ERP) sign of the brain's evaluation of novelty. Neuroscience and Biobehavioral Reviews 25: 355–373.
47. Foulke E, Sticht TG (1969) Review of Research on Intelligibility and Comprehension of Accelerated Speech. Psychological Bulletin 72: 50–&.
48. Versfeld NJ, Dreschler WA (2002) The relationship between the intelligibility of time-compressed speech and speech in noise in young and elderly listeners. Journal of the Acoustical Society of America 111: 401–408.
49. Pisoni DB (1993) Long-Term-Memory in Speech-Perception - Some New Findings on Talker Variability, Speaking Rate and Perceptual-Learning. Speech Communication 13: 109–125.
50. Vaughan N, Storzbach D, Furukawa I (2006) Sequencing versus nonsequencing working memory in understanding of rapid speech by older listeners. Journal of the American Academy of Audiology 17: 506–518.
51. Budhiraja A, Chen LJ, Lee C (2007) A survey of numerical methods for nonlinear filtering problems. Physica D-Nonlinear Phenomena 230: 27–36.
52. Judd K, Smith LA (2004) Indistinguishable states II - The imperfect model scenario. Physica D-Nonlinear Phenomena 196: 224–242.
53. Nowak MA, Krakauer DC, Dress A (1999) An error limit for the evolution of language. ProcBiolSci 266: 2131–2136.
54. Kumar S, Stephan KE, Warren JD, Friston KJ, Griffiths TD (2007) Hierarchical processing of auditory objects in humans. PLoSComputBiol 3: e100.
55. Boemio A, Fromm S, Braun A, Poeppel D (2005) Hierarchical and asymmetric temporal sensitivity in human auditory cortices. NatNeurosci 8: 389–395.
56. Denham SL, Winkler I (2006) The role of predictive models in the formation of auditory streams. Journal of Physiology-Paris 100: 154–170.
57. Badre D (2008) Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends Cogn Sci.
58. Botvinick MM (2008) Hierarchical models of behavior and prefrontal function. Trends Cogn Sci.
59. Koechlin E, Summerfield C (2007) An information theoretical approach to prefrontal executive function. Trends Cogn Sci 11: 229–235.
60. Todorov E, Li W, Pan X (2005) From task parameters to motor synergies: A hierarchical framework for approximately-optimal control of redundant manipulators. JRobotSyst 22: 691–710.
61. Chechik G, Anderson MJ, Bar-Yosef O, Young ED, Tishby N, et al. (2006) Reduction of information redundancy in the ascending auditory pathway. Neuron 51: 359–368.
62. Nelken I (2008) Processing of complex sounds in the auditory system. Current Opinion in Neurobiology 18: 413–417.
63. Nunez PL, Silberstein RB (2000) On the relationship of synaptic activity to macroscopic measurements: Does co-registration of EEG with fMRI make sense? Brain Topography 13: 79–96.
64. Garrido MI, Friston KJ, Kiebel SJ, Stephan KE, Baldeweg T, et al. (2008) The functional anatomy of the MMN: A DCM study of the roving paradigm. Neuroimage 42: 936–944.
65. Maess B, Herrmann CS, Hahne A, Nakamura A, Friederici AD (2006) Localizing the distributed language network responsible for the N400 measured by MEG during auditory sentence processing. Brain Research 1096: 163–172.
66. Van Petten C, Luka BJ (2006) Neural localization of semantic context effects in electromagnetic and hemodynamic studies. Brain and Language 97: 279–293.
67. Marslen-Wilson WD, Bouma H, Bouwhuis D (1984) Function and process in spoken word recognition. Attention and Performance X: Control of Language Processes 125–150, Hillsdale, N.J.: Erlbaum.
68. McClelland JL, Elman JL (1986) The Trace Model of Speech-Perception. Cognitive Psychology 18: 1–86.
69. Norris D (1994) Shortlist - A Connectionist Model of Continuous Speech Recognition. Cognition 52: 189–234.
70. Norris D, McQueen JM (2008) Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review 115: 357–395.
71. Botvinick MM, Plaut DC (2006) Short-term memory for serial order: a recurrent neural network model. PsycholRev 113: 201–233.
72. Broome BM, Jayaraman V, Laurent G (2006) Encoding and decoding of overlapping odor sequences. Neuron 51: 467–482.
73. Davis MH, Johnsrude IS (2007) Hearing speech sounds: top-down influences on the interface between audition and speech perception. HearRes 229: 132–147.
74. Fee MS, Kozhevnikov AA, Hahnloser RH (2004) Neural mechanisms of vocal sequence generation in the songbird. AnnNYAcadSci 1016: 153–170.
75. Ji D, Wilson MA (2008) Firing rate dynamics in the hippocampus induced by trajectory learning. JNeurosci 28: 4679–4689.
76. Koechlin E, Jubault T (2006) Broca's area and the hierarchical organization of human behavior. Neuron 50: 963–974.
77. Kumaran D, Maguire EA (2007) Match mismatch processes underlie human hippocampal responses to associative novelty. JNeurosci 27: 8517–8524.
78. Nadasdy Z, Hirase H, Czurko A, Csicsvari J, Buzsaki G (1999) Replay and time compression of recurring spike sequences in the hippocampus. JNeurosci 19: 9497–9507.
79. Redcay E (2008) The superior temporal sulcus performs a common function for social and speech perception: implications for the emergence of autism. NeurosciBiobehavRev 32: 123–142.
80. Berns GS, Sejnowski TJ (1998) A computational model of how the basal ganglia produce sequences. Journal of Cognitive Neuroscience 10: 108–121.
81. Elman JL (1990) Finding Structure in Time. Cognitive Science 14: 179–211.
82. Jensen O, Lisman JE (1996) Theta/gamma networks with slow NMDA channels learn sequences and encode episodic memory: Role of NMDA channels in recall. Learning & Memory 3: 264–278.

83. Kulvicius T, Porr B, Worgotter F (2007) Chained learning architectures in a simple closed-loop behavioural context. Biological Cybernetics 97: 363–378.
84. Namikawa J, Tani J (2008) A model for learning to segment temporal sequences, utilizing a mixture of RNN experts together with adaptive variance. Neural Netw 21: 1466–1475.
85. Tani J (2003) Learning to generate articulated behavior through the bottom-up and the top-down interaction processes. Neural Netw 16: 11–23.
86. Wyss R, Konig P, Verschure PFMJ (2006) A model of the ventral visual system based on temporal stability and local memory. Plos Biology 4: 836–843.
87. Cleeremans A, McClelland JL (1991) Learning the Structure of Event Sequences. Journal of Experimental Psychology-General 120: 235–253.
88. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computation 9: 1735–1780.
89. Berniker M, Kording K (2008) Estimating the sources of motor errors for adaptation and generalization. Nature Neuroscience 11: 1454–1461.
90. Sakata JT, Brainard MS (2008) Online Contributions of Auditory Feedback to Neural Activity in Avian Song Control Circuitry. Journal of Neuroscience 28: 11378–11390.
91. Tani J (2007) On the interactions between top-down anticipation and bottom-up regression. Front Neurorobotics 1: 2.
92. Yamashita Y, Tani J (2008) Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: A Humanoid Robot Experiment. Plos Computational Biology 4.
93. George D (2008) How the brain might work: A hierarchical and temporal model for learning and recognition [Ph.D.]: Stanford University.