HOW DOES THE BRAIN RECOGNIZE SPEECH – MODELLING USING HIERARCHICAL RECURRENT NEURAL NETWORKS

Stefan J. Kiebel, Burak I. Yildiz

Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany
kiebel@cbs.mpg.de

Abstract: How does the brain recognize speech? In cognitive neuroscience, this question is usually addressed by experiments using neuroimaging methods, e.g. functional magnetic resonance imaging (fMRI) and electroencephalography (EEG). Although there is tremendous progress in better understanding how the human brain recognizes speech, there is actually little progress in elucidating the computational mechanisms of how this is achieved. Here, I present a recently developed computational model which uses recent neurobiological insights from another species, songbirds. Using this computational model, we show that a fusion of two well-established computational approaches, recurrent neural networks and Bayesian filtering, can be used to recognize both birdsong and human speech. The recurrent neural network is based on sequential dynamics as implemented by heteroclinic channels and Hopfield attractor networks. The Bayesian filtering uses a recent formulation which enables online decoding of hierarchical, stochastic, nonlinear dynamical systems. In summary, this model may, on one hand, be an appropriate model for testing quantitative predictions in cognitive neuroscience experiments and, on the other, a novel machine learning tool for artificial speech recognition.

1 Introduction

Speech recognition is a fascinating field where computationally inclined researchers work on developing artificial speech recognition (ASR) algorithms, and cognitive neuroscientists work, in parallel, on speech recognition as performed by human subjects. One may expect, in principle, strong interactions between these two fields. However, as observed from the cognitive and computational neurosciences, this does not seem to be the case. One reason may be that there is no common modelling ground, i.e. cognitive neuroscientists usually employ rather coarse-grained non-mathematical models or concepts about how the brain recognizes speech while ASR researchers have to provide working computational solutions, which are not necessarily considered neurobiologically plausible [1-3].

Here we describe a computational modelling approach which may, in principle, be useful for both cognitive neuroscience and ASR or machine learning. The approach is based on experimental and computational evidence that the brain uses a hierarchy of time scales to perform robust and accurate auditory online recognition [4-8]. Furthermore, to achieve neurobiological plausibility, we use a recently established neural network approach using continuous nonlinear dynamics. Critically, we apply Bayesian filtering to a recurrent neural network to derive update equations which exchange predictions and predictions error messages between neurons [9, 10]. The resulting system shares several key features with speech recognition as done by the brain: it (i) operates online, (ii) predicts its input, (iii) employs multi-scale decoding, and (iv) is generally robust under adverse conditions.

In the following, we will briefly describe the model, and will illustrate the main features of the approach using applications. Finally, we discuss the relevance of this model for auditory speech recognition both in cognitive neuroscience and ASR.

2 Modelling

The overall goal is to develop a model that can track hidden states of a speaker causing the on-going sound waves as sensory input to the recognizing system. To do this, we use hierarchically structured, continuous, stochastic nonlinear dynamical systems as a generative model and use Bayesian filtering to infer about the hidden states given the sensory input [11].

As we will show below, the recognizing system will compute continuous predictions and prediction errors in an online fashion. We will demonstrate that the prediction error can be used to classify words with high performance.

How can one derive a neurobiologically plausible model of speech recognition? For obvious reasons, experiments using human subjects are usually performed non-invasively and experimental findings are too coarse-grained both in space and time to elucidate the computational mechanism at a microscopic neuronal scale. Interestingly, the songbird brain has to solve a similar task, i.e. to decode a song of a complex time-frequency structure and infer hidden variables like the sequence of syllables and motifs, or even the fitness of the singing bird. In birds, the output and neuronal responses of the song generation system can be measured precisely at a microscopic level and this has resulted in a considerable body of experimental findings. Therefore, we used some of the key findings to assemble a complete model of birdsong generation and use it as the basis for constructing a potentially neurobiologically plausible, artificial recognition system based on state-of-the-art Bayesian inference techniques. This is described in detail in [11] and rehearsed here for completeness. We then use the birdsong model with some adaptations for speech recognition.

2.1 A generative model of birdsong

A birdsong consists of small units called notes (analogous to phonetic units in speech) which can be grouped together to form syllables [12]. A combination of identical or different syllables forms motifs. This hierarchical structure of song units is produced by two highly specialized song pathways [13]. In the motor pathway, the forebrain nucleus HVC includes specific neurons called HVC (RA) that project to nucleus RA. RA neurons innervate the vocal and respiratory nuclei to produce vocal output. Our modelling approach is based on the following key experimental observations: During birdsong generation, HVC (RA) neurons fire sequentially at temporally precise moments where each element of this sequence fires only once during the song to control a group of RA neurons [14-16]. This suggests that bursting HVC (RA) neurons select and drive the activity of subsets of RA neurons [14]. In particular, each RA neuron can be driven by more than one HVC (RA) neuron, see Figure 1.

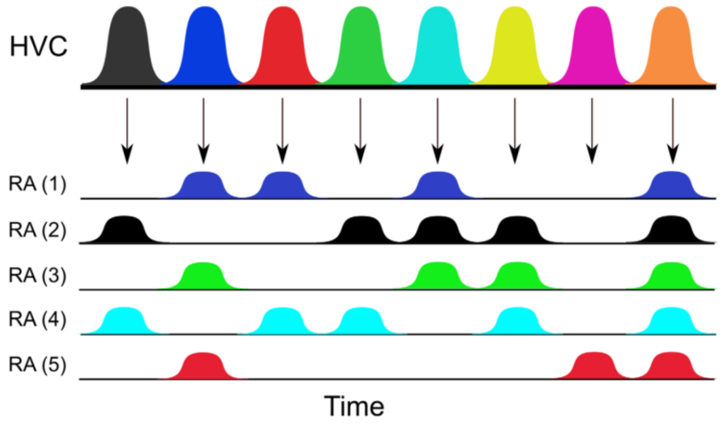

Figure 1: The scheme of HVC and RA dynamics. Five RA ensembles are controlled by eight sequentially activated HVC (RA) ensembles. The horizontal axis denotes time and the arrows describe the specific HVC ensemble that activates the corresponding RA ensembles. The color scheme matches the dynamics shown in Figure 2. Adapted from [11].

How can one model such a mechanism? There have been several approaches to model the sequential activation of HVC (RA) neurons using single neuron models, e.g. [17]. Here, we follow an alternative way by capturing the neuronal mass activity using firing rate models, i.e. we consider model neurons that can be thought of as the synchronized firing activity of an ensemble of neurons. This is motivated by experimental evidence suggesting that there are about 200 co-active HVC (RA) neurons at a specific time during song generation [14]. One of the well established ways for modelling the sequential activation of neuronal ensembles is the winnerless competition using Lotka-Volterra type dynamics [6, 18]. This approach aims at modelling activity at a mesoscopic level, e.g. activity that may be expressed in local field potentials.

When HVC (RA) ensembles undergo sequential activations, the RA level is driven from one attractor to the next. Such networks with attractor dynamics (Hopfield networks) can encode a large number of potential attractors because the forcing input from the HVC level effectively recombines subsets of RA ensembles in distinct assemblies.

At the lowest level, we map the dynamical RA states onto motor neurons. To do this, we compute linear combinations of oscillators at different frequencies which represent the effect of currently active RA ensembles and create dynamical control signals for a model of the vocal organ, the syrinx [19]. This mathematical model of the syrinx has been used previously to model several birdsongs [20].

In summary, the present three-level hierarchical model generates sequences at its top (HVC) level, which are transformed into sequences of multi-dimensional attractors at the RA level. The model consists of hierarchically coupled stochastic, nonlinear, differential equations which are described in detail in [11]. The output of this system are vocal control signals which when used in the syrinx model generate a realistically sounding birdsong sonogram.

2.2 Online Bayesian Recognition

The inference is based on hierarchical message passing and implements a predictive coding scheme for dynamics. All the update equations of the recognition system (to reconstruct the hidden states) consist of differential equations (as in the generation model, see 2.1) and therefore may be implemented by neuronal populations and their network interactions via forward, backward and lateral connections [21]. In general, we assume that listening birds have internal models for the songs they have learned before and the generative model of the heard songs should fit to this internal model. Using this concept, we model optimal recognition using Bayesian inference for hierarchical, nonlinear dynamical systems [21]. For the sensory input, we assume that the vocal control signal given the sound wave, can be readily extracted by the listening bird (agent) from the spectrotemporal dynamics. Given this vocal control signal, we infer the hidden, spatiotemporal RA dynamics and the sequential HVC (RA) dynamics in an online fashion. The proposed Bayesian inference scheme provides, under some assumptions, optimal inference to decode the RA and HVC (RA) dynamics, i.e. to recognize the hidden messages embedded into the vocal control signal. The mathematical description is omitted here and can be conceptualized as follows: At each time step *t*, the recognition system receives sensory input, here the current amplitudes of the vocal control dynamics. Like the generative model, the recognition system has three levels as well. Each of these three levels consists of interacting neuronal populations, which encode predictions, i.e. expectations, about how their internal dynamics will evolve during a song. At the same time, each level receives input from the subordinate level. For the first level, this is the sensory input, which is compared with the internal prediction. The prediction error is forwarded to the second level, where again predictions are used to generate prediction errors, which are forwarded to the third level. Critically, each level adjusts its internal predictions to minimize its prediction error weighted by the prior precision of the internal prediction. At each level, the updated predictions are sent to the subordinate levels to guide their internal predictions by higher level predictions. In summary, each level minimizes its prediction error by a fusion of internal dynamics with top-down (predictions) and bottom-up (prediction error) messages. The overall result is that a listening bird fuses its dynamic and hierarchically arranged expectations about a song with the actual sensory input. Importantly, due to this dynamic fusion, the recognition is robust against deviations from its expectations by explaining away errors of the singing bird by internal precision-weighted prediction error. The derivation of the update equations to achieve Bayes-optimal online recognition solutions is non-trivial, see [21].

2.3 Extension to speech

How can the model of recognizing birdsong be translated to human speech? In on-going work (paper in preparation), we will show that this is straightforward by removing the songbird-specific, lowest level of the hierarchy described in section 2.1 and replacing the vocal control signal input by cochleagrams computed directly from sound waves. In addition, we have used the Bayesian filtering framework [21] to implement learning of cochleagrams. The weights to be learnt are the connection weights from the third to the second level (from the stable heteroclinic channel to the Hopfield network). We found that learning is easy. The intuition behind this finding is that the agent has an internal model which assumes (due to the stable heteroclinic channels) that the sensory input consists of sequences. Since this is the case for speech signals as expressed in a cochleagram, the only thing to learn is how each element of a sequence is expressed in the input, which is a relatively simple learning task.

To provide for a proof-of-concept, we used an established benchmark test (taken from the Texas Instrument TI-46 isolated speech database, available from www.ldc.upenn.edu). There were five (female) speakers, speaking the digits from zero to nine, ten times, providing for, in total, for 500 isolated words. We computed from these words a cochleagram, and subdivided the frequency bands in six bands over which we averaged across frequency. This reduction of the data was done to keep computation time low.

Classification of words proceeded as follows. For each digit word to be learnt, a specific agent was learned across different speakers (see below). Using a cross-validation approach, we used a subset of the ten trials for each speaker and digit as training data, while the classification was performed on the previously unseen remaining trials. As a classification measure, we simply used the accumulated prediction error, over time and neurons and used the argmin over all ten agents.

3 Results

3.1 Recognition of birdsong

We found that the online recognition of simulated birdsong using the Bayesian filtering approach described above provided for rapid and accurate recognition. In Fig. 2, we show the recognition result for the high signal-to-noise ratio case ('ideal communication'). It can be seen that all hidden states, at all three levels are reconstructed accurately. Note that this reconstruction is online, i.e. the recognition agent reconstructs on the fly. This simulation provides an example of how a bird may be able to extract multi-scale information from fast-varying sensory input. In [11], we describe further simulations, which show that the proposed online recognition is robust against noise and unexpected perturbations. Furthermore, we use simulations to address specific neuroscience findings, e.g. the cooling of HVC which has been found to slow down the song generation.

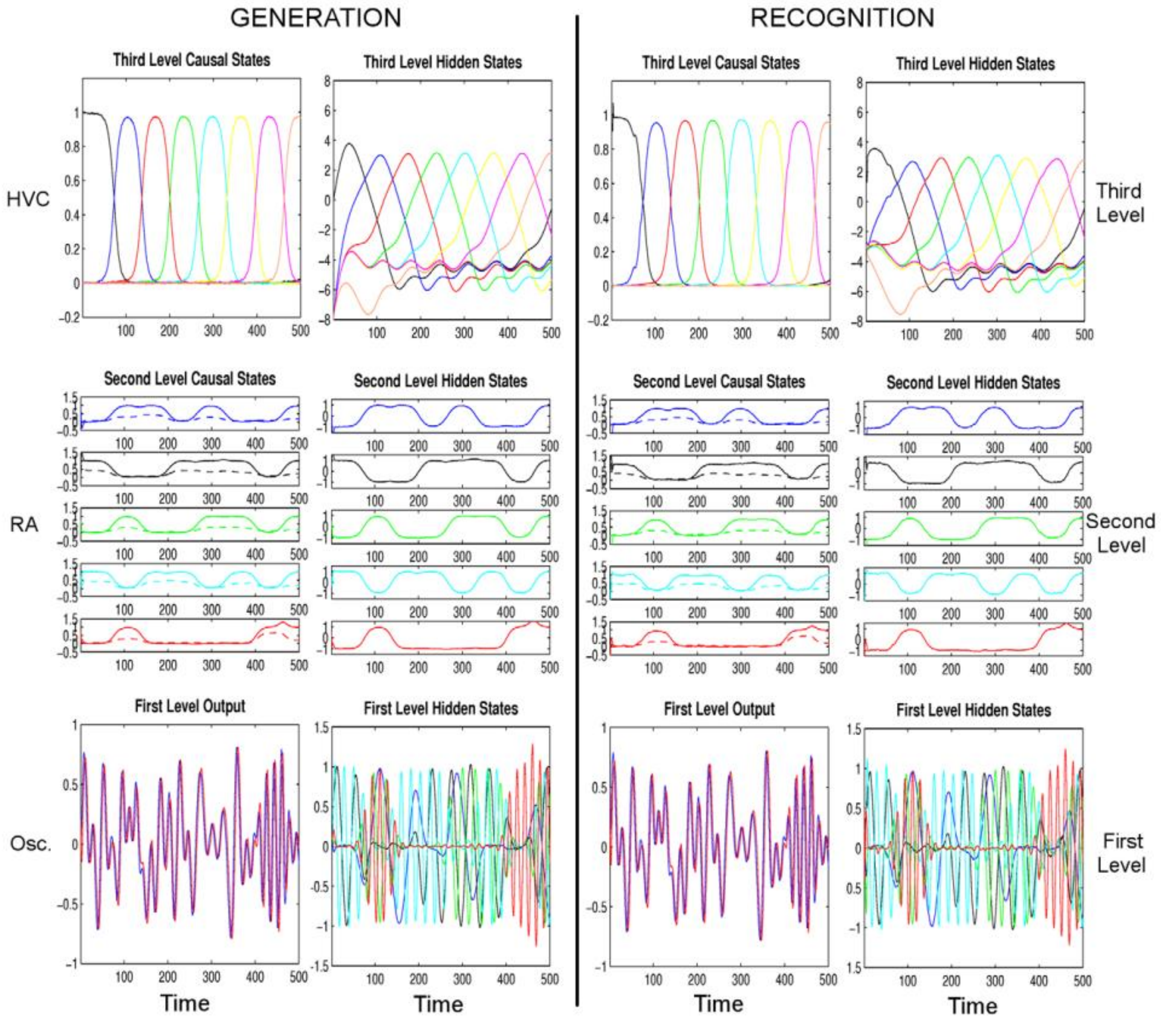

Figure 2: Generated and recognizing dynamics for simulated birdsong. (Left) The causal states are shown on the left and hidden states on the right with arbitrary units both in time and neuronal activation. There are three levels: A) HVC (third) level, B) RA (second) level and C) Oscillator (first) level. At the HVC level, there are eight HVC ensembles (each represented with a different colour) which are activated for a short amount of time to control the dynamics of the five RA ensembles, see also Figure 1. At the second level (left column), the solid and dashed lines represent causal states, and the dashed lines represent the hidden states, also shown in the right column. The first level models the vocal control signals. The signals shown in the left column are the output dynamics which control the syrinx to obtain synthetic birdsong and are the input to the recognition. (Right) The format of the recognizing dynamics are the same as for the generated dynamics on the left. The recognition scheme receives only the output of the first level (bottom left) and reconstructs the hidden states at all levels using the online Bayesian inference scheme. It can be seen that the reconstruction is successful as there are only tiny deviations between the true (left) and the reconstructed (right) dynamics. Adapted from [11].

3.2 Learning and recognition of speech

As described above, we translated the birdsong model to human speech cochleagrams and used an isolated word benchmark test to show that the model performs well in such a benchmark (paper in preparation). We found that, using the original sound waves as a starting point, and after learning, a collection of ten speech recognition agents could classify 98.4% of the test words, using a cross-validation procedure. When adding white noise to the sound waves, the performance was found to drop slightly but stayed reasonably high.

	Original	**Noise 30db**	**Noise 20db**	**Noise 10db**
Accuracy	**98.4 %**	96.4 %	95 %	88.8 %

Table 1: Recognition rates using the digit words from the Texas Instrument TI-46 isolated speech database. Each entry lists the recognition rate determined using a cross-validation procedure. To achieve different signal-to-noise ratios, we added white noise with different variances to the original sound wave data before computing the cochleagrams (see text).

4 Discussion

We have described a novel artificial recognition algorithm for both birdsong and speech. The algorithm is based on the application of Bayesian filtering to hierarchically structured, stochastic, recurrent neural networks [10, 11]. The results show that fast sensory input dynamics can be tracked and mapped, in an online fashion, to hidden states, see Figure 2. These reconstructed hidden state dynamics of the recognition agent represent the states of the generating agent which caused the sensory input. In other words, the recognition agent can infer from the sensory input on the hidden states of the sender in an online fashion, which is an effective way of communication. Remarkably, this scheme works across multiple levels and may be the basis for the multi-scale recognition performance of the brain. We also found the scheme to be robust against noise and perturbations.

When applying this scheme to an isolated word data set, we found that we can learn and recognize at high performance, see Table 1. We expect that we can improve these recognition rates further, because we, for computational reasons, had to reduce the cochleagram to just six frequency averages (out of eighty) for each time point. It remains to be shown whether the present recognition system can be scaled up to larger speech data bases and can perform continuous speech recognition as required for artificial speech recognition.

Interestingly, the concept underlying this recognition scheme is very similar to the conventional hidden Markov model (HMM): It is assumed that speech is a sequence like 'beads on a string' [22]. However, there are critical differences, which help to overcome some difficulties with the HMM in the application to speech: (i) the present recognition system uses nonlinear dynamical system as a basis. This means that parameters do not encode directly transition probabilities between discrete states but rather describe a transient sequence of saddle points. This makes the model parameterization much more parsimonious than as in HMMs because the generating dynamics only need to describe the sequence of saddle points but not exactly how to get from one to the next: The passage between saddle points of the stable heteroclinic channels (HVC level in Fig. 1 and 2) is implicitly described by the nonlinear dynamics. (ii) For the same reason it is not necessary to split speech data in fixed-length or variable-length data chunks because the model is continuous and appropriate for continuous data like speech sound waves. For example, with the isolated word data set (section 3.2), we arbitrarily chose reference points between saddle points and specific time points in the cochleagram but did not segment the data themselves. (iii) Vocal tract movements have slow components which modulate speech dynamics across multiple time scales. Although we have not modelled these effects here, nonlinear dynamical systems may be ideal to model such effects because modulation of on-going dynamics by previous dynamical events is naturally expressed in nonlinear dynamical systems as a perturbation of states. Specifically, the induced changes may be explained away by the on-going dynamics but, in principle, do the cross-temporal interactions do need to be modelled by additional parameters as in HMMs.

Acknowledgments

The Max Planck Institute funded this work. We thank Katharina von Kriegstein for helpful discussions.

References

1. Bilmes, J.A., *What HMMs can do.* Ieice Transactions on Information and Systems, 2006. **E89d**(3): p. 869-891.
2. Deng, L., D. Yu, and A. Acero, *Structured speech modeling.* Ieee Transactions on Audio Speech and Language Processing, 2006. **14**(5): p. 1492-1504.
3. O'Shaughnessy, D., *Invited paper: Automatic speech recognition: History, methods and challenges.* Pattern Recognition, 2008. **41**(10): p. 2965-2979.
4. Kiebel, S.J., J. Daunizeau, and K.J. Friston, *A Hierarchy of Time-Scales and the Brain.* Plos Computational Biology, 2008. **4**(11).
5. Kiebel, S.J., J. Daunizeau, and K.J. Friston, *Perception and hierarchical dynamics.* Front Neuroinform, 2009. **3**: p. 20.
6. Kiebel, S.J., et al., *Recognizing Sequences of Sequences.* Plos Computational Biology, 2009. **5**(8).
7. Lerner, Y., et al., *Topographic mapping of a hierarchy of temporal receptive windows using a narrated story.* J Neurosci, 2011. **31**(8): p. 2906-15.
8. Poeppel, D., W.J. Idsardi, and W.V. van, *Speech perception at the interface of neurobiology and linguistics.* Philos.Trans.R.Soc.Lond B Biol.Sci., 2008. **363**(1493): p. 1071-1086.
9. Friston, K. and S. Kiebel, *Predictive coding under the free-energy principle.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2009. **364**(1521): p. 1211-1221.
10. Bitzer, S. and S.J. Kiebel, *Recognizing recurrent neural networks (rRNN): Bayesian inference for recurrent neural networks.* Biol Cybern, 2012. **106**(4-5): p. 201-17.
11. Yildiz, I.B. and S.J. Kiebel, *A Hierarchical Neuronal Model for Generation and Online Recognition of Birdsongs.* Plos Computational Biology, 2011. **7**(12).
12. Doupe, A.J. and P.K. Kuhl, *Birdsong and human speech: common themes and mechanisms.* Annu Rev Neurosci, 1999. **22**: p. 567-631.
13. Bolhuis, J.J. and M. Gahr, *Neural mechanisms of birdsong memory.* Nat Rev Neurosci, 2006. **7**(5): p. 347-57.
14. Fee, M.S., A.A. Kozhevnikov, and R.H. Hahnloser, *Neural mechanisms of vocal sequence generation in the songbird.* Ann N Y Acad Sci, 2004. **1016**: p. 153-70.
15. Hahnloser, R.H., A.A. Kozhevnikov, and M.S. Fee, *An ultra-sparse code underlies the generation of neural sequences in a songbird.* Nature, 2002. **419**(6902): p. 65-70.
16. Yu, A.C. and D. Margoliash, *Temporal hierarchical control of singing in birds.* Science, 1996. **273**(5283): p. 1871-5.
17. Li, M. and H. Greenside, *Stable propagation of a burst through a one-dimensional homogeneous excitatory chain model of songbird nucleus HVC.* Phys Rev E Stat

Nonlin Soft Matter Phys, 2006. **74**(1 Pt 1): p. 011918.

18. Rabinovich, M., et al., *Dynamical encoding by networks of competing neuron groups: Winnerless competition.* Physical Review Letters, 2001. **8706**(6).
19. Laje, R., T.J. Gardner, and G.B. Mindlin, *Neuromuscular control of vocalizations in birdsong: a model.* Phys Rev E Stat Nonlin Soft Matter Phys, 2002. **65**(5 Pt 1): p. 051921.
20. Laje, R. and G.B. Mindlin, *Diversity within a birdsong.* Phys Rev Lett, 2002. **89**(28 Pt 1): p. 288102.
21. Friston, K., *Hierarchical models in the brain.* PLoS Comput.Biol., 2008. **4**(11): p. e1000211.
22. Ostendorf, M., *Moving beyond the 'beads-on-a-string' model of speech.* Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, 1999. **1**: p. 5.