

Research report

# Identification of prosodic attitudes by a temporal recurrent network

Jean-Marc Blanc, Peter Ford Dominey\*

*Institut des Sciences Cognitives, UMR 5015 CNRS, University Claude Bernard Lyon 1, 67 Boulevard Pinel, 69675 Bron Cedex, France*

Accepted 25 June 2003

## Abstract

Human speakers modulate the fundamental frequency (F0) of their utterances in order to express different ‘prosodic’ attitudes such as surprise or curiosity. How are these prosodic attitudes then decoded? The current research addresses the issue of how the temporal structure of F0 can be used in order to discriminate between prosodic attitudes in natural language using a temporal recurrent neural network (TRN) that was initially developed to simulate the neurophysiology of the primate frontostriatal system. In the TRN, a recurrent network of leaky integrator neurons encodes a continuous trajectory of internal states that characterizes the input sequence. The input to the model is a population coding of the continuous, time-varying values of the fundamental frequency (F0) of natural language sentences. We expose the model to an experiment based on one in which human subjects were required to discriminate between different prosodic attitudes (surprise, exclamation, question, etc.). After training, the model discriminates between six prosodic attitudes in new sentences at 82.52% correct, compared to 72.8% correct for human subjects. These results reveal (1) that F0 provides relevant information for prosodic attitude discrimination, and (2) that the TRN demonstrates a categorical sensitivity to this information that can be used for classifying new sentences.

© 2003 Elsevier B.V. All rights reserved.

*Theme:* Neural basis of behavior

*Topic:* Learning and memory: systems and functions

*Keywords:* Temporal sequence learning; Neural network; Frontostriatal system; Recurrent network; Prosodic attitude; Speech

## 1. Introduction

Essentially all of human cognition takes place in time, and thus in a sequential context. This can be seen in examples such as in game playing, problem solving, sensorimotor control and the perception and production of music and language. Recent work in the study of infant perception of language has argued that sensitivity to temporal structure of speech is present at birth [17], and to its serial structure at or before 8 months of age [21]. We have recently demonstrated that the temporal recurrent network (TRN), a recurrent network of leaky integrator neurons, demonstrates this sensitivity to serial and temporal structure of language [6]. The model discriminated between languages from different rhythm classes, based on the temporal structure of consonant–vowel sequences into

which the sentences had been re-coded. The current study extends the processing of the TRN from discrete consonant–vowel coding to allow processing of the continuous, time varying values of the fundamental frequency signal in a population code in a prosodic attitude discrimination task.

Fundamental frequency or F0 is the component of language that allows a sentence such as ‘You went to the bank’ to be uttered and interpreted as a statement of fact, or a question, depending on whether the fundamental frequency falls, or rises at the sentence’s end. The frequency contour and the rhythm of the sentences can be used to determine the speaker’s intended prosodic attitudes [16]. These two dimensions are part of prosody, which is also expressed by modulation of amplitude and related spectral changes. In the current study, only the F0 component will be considered.

The working hypothesis of this research is that a network of recurrently connected leaky integrator neurons should be capable of representing the continuous temporal

\*Corresponding author. Tel.: +33-4-3791-1265; fax: +33-4-3791-1210.

E-mail address: [dominey@isc.cnrs.fr](mailto:dominey@isc.cnrs.fr) (P.F. Dominey).

structure of the fundamental frequency in spoken language in order to perform a prosodic discrimination task. Recurrent networks [5–11,18] are inherently well suited for sequence learning. In these systems, information about previous events that is necessary to predict subsequent events is maintained as an internal state via the recurrent connections. We recently described a model of sensory-motor sequence learning based on the primate frontostriatal system, in which the prefrontal cortex (PFC) is modeled as a recurrent network that encodes a continuously evolving sequence of internal states, and the striatum (caudate nucleus, CD) is modeled as an associative memory structure that binds internal states encoded in PFC to their corresponding motor outputs [9,10]. While this model falls into the general category of recurrent networks, it has certain novel features regarding the processing of temporal structure that will shortly become apparent, that makes it particularly well suited to address the task of prosodic attitude discrimination based on temporal F0 structure.

## 2. The temporal recurrent network (TRN) model

The model architecture that will be used in the rest of this paper, presented in Fig. 1, relies on a recurrent network to represent sequential state. The general principle is that the neuron-like elements in the recurrent network generate a succession of distinct patterns of activation during the processing of an input sequence. The resulting pattern at the end of the presentation of a sequence thus characterizes the sequence and can be used to discriminate between sequences from different temporal structure categories.

### 2.1. Architecture of the network

In a pioneering study of sensorimotor sequence learning, Barone and Joseph [1] demonstrated that neurons in the dorso-lateral prefrontal cortex (PFC) of macaque monkeys

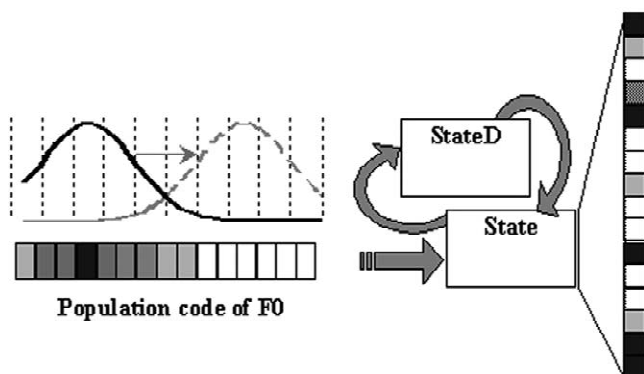


Fig. 1. Architecture of the temporal recurrent network (TRN), and the distributed Gaussian representation of the population-coded fundamental frequency (F0) inputs.

encode both the spatial (retinotopic) location of sequence elements, and their context or rank in the sequence. This suggested that recurrent connections in the cortex could allow neural activity related to previous sequence elements to influence the coding of the current element, thus yielding the observed sequence context encoding. In this framework, PFC would act as a dynamical system, whose activity state would be influenced both by the serial order of sensory inputs, and their temporal structure of durations and delays.

We exploited this idea in a model of sensorimotor sequence learning in which PFC was modeled as a network of recurrently connected leaky integrator neurons [9]. Sequence state was thus encoded in PFC, and it influenced motor output for sequence execution via learning-related modifications in plastic synapses between cortex and striatum. Fig. 2 illustrates the dynamic temporal behavior of this system in a temporal discrimination task [7]. In the current study, we retain the essential component of the recurrent state network, and replace the learning by a functionally equivalent but more efficient method described below. The network consists of three arrays of leaky integrator neurons. The input layer contains 15 units, and two layers of  $5 \times 5$  units (State and State<sub>D</sub>) make up the recurrent network component [6,7]. The response latency of these neuron-like units is a function of their input intensity and their time constant. The ensemble of units are updated every simulation time step or STS, which corresponds to 5 ms of real time.

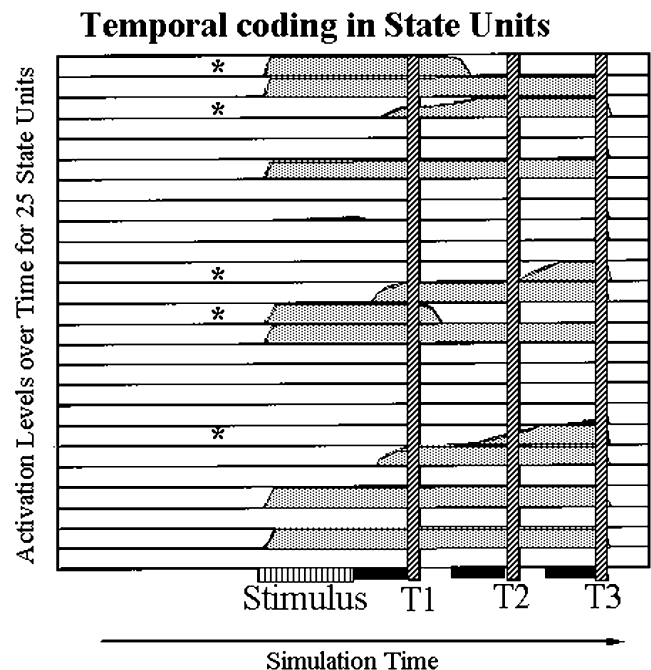


Fig. 2. Temporal coding in State units during a temporal discrimination task. Modification in State activity is seen to explicitly encode the passage of time after the stimulus. Each row represents the activity of the 25 State units over time (from Ref. [7]).

The recurrent network architecture described here is similar to previous recurrent models with several important differences. Firstly, there is no learning in the recurrent connections (i.e. those that project from State<sub>D</sub> to State). Instead the recurrent connections are initialized to random values between  $-0.5$  and  $+0.5$ , providing the network with an inherent dynamical system behavior, and sensitivity to serial and temporal structure of input sequences. In this manner, the space and time complexity of back propagation through time, and recurrent back propagation are avoided [18]. Instead, learning is based on an association between activation vectors generated by sequences in the layer State<sub>D</sub> with appropriate output responses, described below. In this context, an integral part of this model that was originally developed to simulate primate behavioral electrophysiology experiments [9] is the experimenter’s ability to specify the temporal structure of event sequences, making it ideal for studying the effects of serial and temporal structure [7–9]. A related and important difference is that the computing elements are leaky integrators with sigmoid output functions, and thus have a temporal build-up and reduction of activity that contributes to the overall sensitivity to temporal structure.

Sequences are presented to the network by successive activation of units in the input layer which project these inputs to the recurrent network made up of layers State and State<sub>D</sub> defined in Eqs. (1) and (2). In the equations that follow, the function  $f(x)$  generates a non-linear (sigmoid) output of  $x$

$$s_i(t + \Delta t) = \left(1 - \frac{\Delta t}{\tau}\right) s_i(t) + \frac{\Delta t}{\tau} \left( \sum_{j=1}^n w_{ij}^{IS} \text{Input}_j(t) + \sum_{j=1}^n w_{ij}^{SS} \text{State}_{Dj}(t) \right) \quad (1.1)$$

$$\text{State} = f(s(t)). \quad (1.2)$$

In (1.1) the leaky integrator,  $s()$ , corresponding to the membrane potential or internal activation of State is described. In (1.2) the output activity level of State is generated as a sigmoid function,  $f()$ , of  $s(t)$ . The term  $t$  is the time,  $\Delta t$  is the simulation time step,  $\tau$  is the leaky integrator time constant. As  $\tau$  increases with respect to  $\Delta t$ , the charge and discharge times for the leaky integrator increase. In the simulations,  $\Delta t$  is 5 ms. For Eqs. (1)–(4), the time constants are 10 ms, except for Eq. (2.1), which has five time constants that are 100, 600, 1100, 1600 and 2100 ms.

The connections  $w^{IS}$  and  $w^{SS}$  define, respectively, the projections from units in Input and State<sub>D</sub> to State. These connections are one-to-all, and are mixed excitatory and inhibitory, and do not change with learning. This mix of excitatory and inhibitory connections ensures that the State network does not become saturated by excitatory inputs,

and also provides a source of diversity in coding the conjunctions and disjunctions of input and previous state information. The  $n$  in the summation terms is 25, corresponding to the linearized size of the  $5 \times 5$  layers.

Recurrent input to State originates from the layer State<sub>D</sub>. State<sub>D</sub> (Eqs. (2.1) and (2.2)) receives input from State, and its 25 leaky integrator neurons have a distribution of time constants from 20 to 420 simulation time steps (100 to 2100 ms), while State units have time constants of two simulation time steps (10 ms). This distribution of time constants in State<sub>D</sub> yields a range of temporal sensitivity similar to that provided by using a distribution of temporal delays [10]. While these time constants are unnaturally high for single neurons, they are consistent with the time constants of locally interconnected assemblies of neurons. Additionally, we have demonstrated that the network is robust to changes in these time constants. When a different range of values was used (from 10 to 1010 ms), this distribution still left the performance intact [10]

$$sd_i(t + \Delta t) = \left(1 - \frac{\Delta t}{\tau}\right) sd_i(t) + \frac{\Delta t}{\tau} (\text{State}_i(t)) \quad (2.1)$$

$$\text{State}_{Dj} = f(sd_j(t)). \quad (2.2)$$

Part of the novelty of the TRN is the use of this dynamic recurrent network of leaky integrator neurons with pre-set recurrent connections. As illustrated in Fig. 2, this provides the system with an inherent dynamic behavior that yields sensitivity to temporal structure.

## 2.2. Coding of fundamental frequency

The fundamental frequency (F0) describes the melodic contour of a sentence. For example, in the sentence ‘You went to the bank’, the final word ‘bank’ can have different temporal F0 profiles. If the F0 profile rises with respect to the rest of the words, then the sentence will be interpreted as a question. If it stays at the same level, it will be interpreted as a simple declaration. In the corpus of sentences that were tested [16], the tone (F0) is expressed in tenths of quarter tones. The values were normalized, and corresponded to the difference between the current frequency of the speaker with his average frequency. Three values represent F0 at 10%, 50% and 90% of the duration for each phoneme. From these data the fundamental frequency was described as a linear interpolation across time. This normalized F0 trajectory was presented as input to the network.

To represent F0, a Gaussian distribution (Eq. (3)) is used in a population of Input neurons where each neuron represents a specific F0 value. Based on this coding, the more a given neuron is activated, the more it represents the fundamental frequency, and similar F0 frequencies will have similar representations over the population. Alternatively if each neuron discretely characterized a range of frequency, there would exist some neighboring frequencies

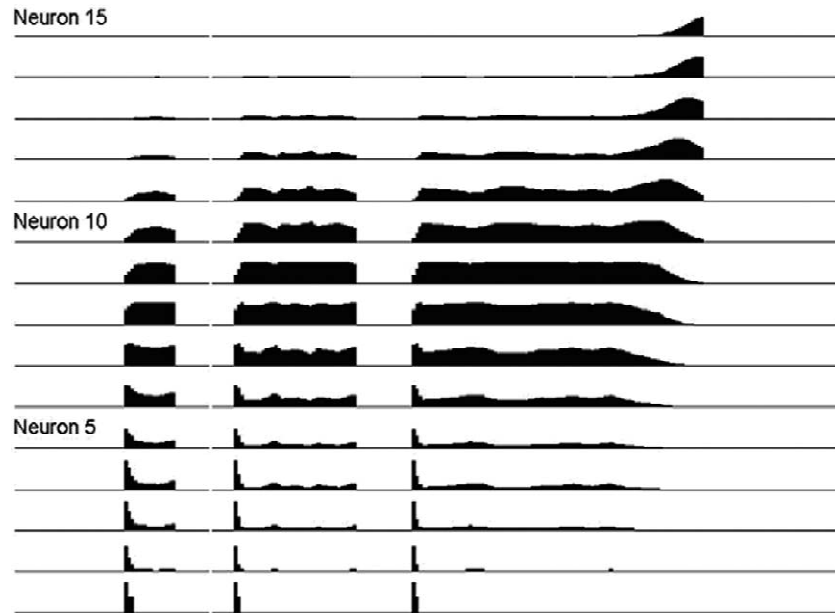


Fig. 3. Representation F0 input structure, represented in the 15 Input neurons, for a sentence expressed with a question attitude.

whose representations would be totally different, corresponding to two different neurons. This distributed Gaussian coding is schematically represented in Fig. 3

$$\text{Activation}(i, f_0) = \text{Max}_{\text{activation}} \sqrt{2\pi} \times \sigma \times \text{NormalLaw}(\text{step}(i + 1/2) + \text{Min}_{f_0}, f_0, \sigma). \quad (3)$$

The Gaussian activation function is described in Eq. (3), where:  $\text{step} = (\text{Max}_{f_0} - \text{Min}_{f_0})/nb$ ,  $nb$  represents the number of neurons used for the coding;  $\text{Max}_{f_0}$  is the **maximum** reached by the fundamental frequency, for all the sentences and attitudes;  $\text{Min}_{f_0}$  is the **minimum** reached by the fundamental frequency, for all the sentences and attitudes;  $\sigma$  represents the standard deviation allowing control of window of the active neurons;  $\text{Max}_{\text{activation}}$  is the value of the most active neuron of the layer Input (here, 70%);  $f_0$  is the current value of the fundamental frequency;  $i$  is the number of the current neuron.

For our experiment a population of 50 TRNs are trained with a group of 60 six-syllable sentences. Each of these sentences was expressed with the six prosodic attitudes by a unique orator. The TRN was then tested with 60 new sentences, each expressed with the same six attitudes, by the same speaker [16].

### 2.3. Training and evaluation of network performance

In this study the performance of interest is the ability of the network to categorize F0 sequences. In order to achieve this, we exploit the encoding of a given sequence in the recurrent State network after the presentation of the

sequence to the network. This snapshot of activity in  $\text{State}_D$  forms a 25-element vector as schematized in Fig. 1. The objective of training is to form associations between (a) State vectors arising from sentences with the same attitudes and (b) the respective categorical response. These acquired associations are then tested as we verify this categorization capability with new sentences. One method to achieve this is with an associative memory as described in Eq. (4). In this case, during supervised training, after an input sentence is processed, the Output neuron corresponding to attitude for that sentence is activated, and Eq. (4) is then applied in order to associate the activity in the State vector, with the activation of the corresponding categorical neuron in Output. Based on this training procedure, State coding of sentences in category  $j$  will become associated with activation of Output neuron  $j$ . In Eq. (4), the matrix  $w$  describes the association between neurons  $i$  in the State vector, and neurons  $j$  in the Output vector, where each neuron  $j$  corresponds to one prosodic attitude category. Thus, after training with a set of input sentences in category  $j$ , the connections in  $w_{ij}$  from State to Output neuron  $j$  encode the average contribution of State neurons to the categorization response  $j$

$$w_{ij}^{\text{SO}}(t + 1) = w_{ij}^{\text{SO}}(t) + R \times \text{State}_{D_i} \times \text{Output}_j. \quad (4)$$

The associative network in Eq. (4) simulated learning-related modification of corticostriatal synapses [9]. This allowed the model to perform sequence discrimination based on serial and temporal structure [6–10].  $\text{State}_D$  vectors that were similar (i.e. that were close together in  $\text{State}_D$  space) would tend to be classified together (i.e. they would activate the same Output neuron  $j$ ), while those that

were different (i.e. distant) would be classified as separable. The problem with this method was that the set of training sequences had to be presented numerous times, each time making small contributions to learning in  $w_{ij}$  in order to avoid interference effects.

In the current study, rather than accumulating the average contribution of a State vector from prosodic category  $j$  to the activation of Output( $j$ ) over multiple learning trials, we take into account the contribution of each State vector in a given category once, by creating for each category, a prototype vector that is the average activation of all State vectors from sentences in that category. The set of thus-formed prototypes corresponds to the set of weights in  $w_{ij}$  that project to the distinct output neurons for each category. This approach is functionally equivalent to the iterative learning performed with Eq. (4), but has significantly reduced computational complexity.

In the following experiment we will be interested in the ability to classify different sequences into their respective prosodic categories. In order to do this, for each category a prototype vector will be constructed as the mean State<sub>D</sub> vector generated from a set of State<sub>D</sub> vectors from sequences in that respective category. Categorization of a new sequence will then be achieved by determining the minimum distance between that State<sub>D</sub> vector and the set of prototype vectors based on Eq. (5)

$$\min_{1 \leq i \leq n} (\|\vec{A} - \vec{P}_{\text{activation}}^i\|_2) \tag{5}$$

where  $\vec{A}$  represents the vector of activation in State<sub>D</sub>.  $\vec{P}_{\text{activation}}^i$  stands for the prototype (mean of activation) of the category  $i$  (from among  $n$  possible categories).

### 3. Experiment: identification of prosodic attitudes

The simulation will be based on the behavioral study in which human subjects categorized sentences based on six prosodic attitudes, reported by Morlec et al. [16]. The six attitudes are defined as follows: (1) *Declaration*; (2) *Simple question*; (3) *Exclamation of surprise*; (4) *Doubt–incredulity* (partial discord with what was previously expressed); (5) *Suspicious irony* (doubt on the assertions of the interlocutor); (6) *Evidence* (profound belief of the speaker in his affirmations).

In Ref. [16] the behavioral task was divided in two sessions. The first session allowed the subjects to become familiar with the six prosodic attitudes. In the second session, 48 new sentences were presented and subjects were required to assign each sentence to one of the six attitudes. Human listeners could identify prosodic attitudes with a global score of 72.8% for 20 subjects. Interestingly, the results presented a discrepancy among attitudes. For example, Declaration and Question are better identified than Exclamation and Suspicious Irony. Table 1 is the confusion matrix that identifies for the group of human

Table 1

Confusion matrix for human subjects. Each value corresponds to the percentage of trials in which a sentence with the row attitude was identified as the column attitude

	DC	QS	EX	DI	SC	EV
DC	<b>88.7</b>	0.3	0	0.6	2.5	7.8
QS	2.2	<b>81.4</b>	4.8	7.1	3.2	1.3
EX	1.6	1.6	<b>72.9</b>	1.5	4.4	4.5
DI	5.7	1.9	8	<b>58.7</b>	17	8.7
SC	9.6	3.9	3.5	25	<b>48.5</b>	9.7
EV	5.8	0.3	2.9	2.9	1.9	<b>86.3</b>

subjects the percentage of identification of each attitude with respect to the others [16].

We now examine the performance of the TRN on this task. Performance will be presented for a population of 50 networks each with a randomly generated set of recurrent connections with values in  $(-0.5-+0.5)$ . A more detailed analysis is then provided for the best five networks from the population. In our simulations, sentences will be encoded based on the fundamental frequency or F0 value that continuously varies over the duration of each given sentence as described above. The population of 50 TRNs are first trained with a group of 60 sentences of six syllables, expressed with the six prosodic attitudes by a single speaker. Validation is then performed with 60 new sentences, each expressed with the same six attitudes, by the same speaker [16].

After training the 50 networks, the average discrimination performance for the trained sentences was 88%. Most importantly, when tested in validation with new, untrained sentences, categorization performance was 70.82%, with the five best networks yielding an average of 82.52% (see Fig. 4). This transfer of performance to the validation set indicates that the activity in the recurrent network provides a representation of the F0 structure that is adequate for the categorical representation of prosodic attitudes.

The confusion matrix for these five networks indicates that the TRN is capable of identifying these prosodic attitudes (see Table 2) with a performance well above the

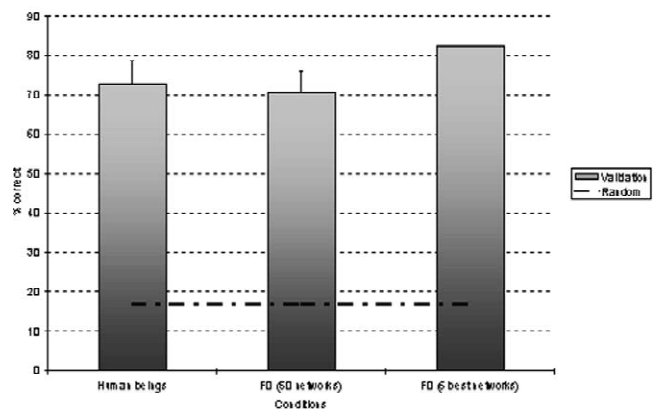


Fig. 4. Performance of human subjects, the 50 networks and the five best networks on the prosodic discrimination validation.

Table 2

Confusion matrix for automatic identification. The results present the mean performance of the five networks, which reach the highest results during the learning period. A global performance of 82.52% was reached

	DC	QS	EX	DI	SC	EV
DC	<b>91.8</b>	1.6	2.6	2.6	1.0	0.3
QS	3.0	<b>90.4</b>	5.3	1.0	0.0	0.3
EX	6.7	1.0	<b>76.2</b>	5.0	0.7	10.4
DI	0.7	0.0	1.3	<b>98.0</b>	0.0	0.0
SC	3.3	3.6	3.0	0.0	<b>85.4</b>	4.6
EV	8.1	2.0	26.2	1.3	3.0	<b>59.4</b>

chance level just as humans did. Interestingly the two modalities Declarative and Question were quite well identified both by the TRN and the human subjects (>80%). However, performance on the other prosodic attitudes (DI, SC, EX and EV) varied between the model and human subjects. This variation is likely due to several issues that are not addressed by the model, including emotional states associated with these attitudes, and different levels of exposure that different subjects may have had to these types of attitudes.

#### 4. Discussion

We previously established that the TRN is able to distinguish between languages from different rhythm-classes, simulating the performance of human infants in this task [6]. This highlighted the sensitivity of the TRN to temporal structure, which is one aspect of the prosodic structure of language. In the current study, the capacity of the model to deal with prosody was significantly extended. In particular, the addition of a capability to represent fundamental frequency allowed the system to identify prosodic attitudes at a level of performance comparable to that of humans. Indeed the TRN seems to be well adapted for the study of prosody because of its relatively simple and realistic treatment of time. That is, the temporal structure of input sequences can be processed and represented in the recurrent network with very low space and time complexity, in contrast to related algorithms including recurrent back-propagation and back-propagation through time which have extensive (and biologically unreasonable) space and time complexity [18].

After a sentence has been presented, the activity pattern in the recurrent network represents a temporal-to-spatial transformation of the F0 structure of the sentence. The observation that these activity patterns preserve prosodic attitude categorical information that can generalize to the categorization of new sentences, is a new and interesting demonstration of how a recurrent network of leaky integrator neurons is sensitive to temporal sequential structure. This represents the first time that a recurrent network, the

TRN, has been used to analyze and categorize the temporal structure of F0 in a prosodic attitude discrimination task. Indeed, while classification of prosodic structure seems to be a natural task for recurrent networks, to date this capability has not been exploited. While we have demonstrated that multiple aspects of temporal structure can be processed in a common network, it appears that human processing of temporal prosodic structure may be functionally lateralized [4,14,15,19], posing a challenge for future simulation studies.

The TRN model was initially developed based on the primate frontostriatal system, including recurrent cortico-cortical connections and modifiable corticostriatal connections. It was developed to learn sensorimotor sequences and to simulate neurophysiological activity during a task performed by non-human primates [9]. The model thus proposed a role for the basal ganglia in sensorimotor sequence learning. Corticostriatal synapses were modified by reward-related dopamine signals in order to bind sequence state representations in cortex to behavioral responses in the striatum. Indeed, numerous computational studies have addressed the sensorimotor learning functions of the basal ganglia (see Ref. [12] for a review) often in the context of the temporal difference (TD) learning algorithm, implemented in the framework of actor-critic models [13,23]. The critic learns to predict the weighted sum of future rewards based on the current sensory input and the actor's policy. Then this prediction is compared to the actual rewards obtained by the actor. Finally the weights of both critic and actor are updated as a function of the error between two adjacent predictions. One recent sequence learning model [12] exploits and explains the observation that the activity of dopamine neurons correspond to the training signal in a TD learning model. In this context, the TD method seems to be well-adapted to discrete sequences [2,3,23], and its application to more continuously structured temporal sequences, such as those representing prosodic attitudes, will be of interest.

The current study contributes to extend the field of sequence learning in recurrent networks in the temporal dimension. In speech, this temporal structure appears to provide important cues for lexical categorization [22], prosodic attitude [16] and language [17,20] identification. The demonstration that the TRN is sensitive to such temporal structure is of potential interest in understanding the underlying neural basis of human sensitivity to these cues.

#### Acknowledgements

Jean-Marc Blanc is supported by a Doctoral Fellowship from the Région Rhone-Alpes. We thank Gérard Bailly for insightful comments and access to the speech data. This work was supported by the Région Rhone-Alpes.

## References

- [1] P. Barone, J.P. Joseph, Prefrontal cortex and spatial sequencing in macaque monkey, *Exp. Brain Res.* 78 (3) (1989) 447–464.
- [2] D.G. Beiser, J.C. Houk, Model of cortical–basal ganglionic processing: encoding the serial order of sensory events, *J. Neurophysiol.* 79 (1998) 3168–3188.
- [3] G.S. Berns, T.J. Sejnowski, A computational model of how the basal ganglia produce sequences, *J. Cogn. Neurosci.* 10 (1) (1998) 108–121.
- [4] T.W. Buchanan, K. Lutz, S. Mirzazade, K. Specht, N.J. Shah, K. Zilles, L. Jancke, Recognition of emotional prosody and verbal components of spoken language: an fMRI study, *Cogn. Brain Res.* 9 (3) (2000) 227–238.
- [5] M.H. Christiansen, R.A.C. Dale, Integrating distributional, prosodic and phonological information in a connectionist model of language acquisition, in: *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, Mahwah, NJ, 2001, pp. 220–225.
- [6] P.F. Dominey, F. Ramus, Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant, *Lang. Cogn. Processes* 15 (1) (2000) 87–127.
- [7] P.F. Dominey, A shared system for learning serial and temporal structure of sensori-motor sequences? Evidence from simulation and human experiments, *Cogn. Brain Res.* 6 (1998) 163–172.
- [8] P.F. Dominey, Influences of temporal organization on transfer in sequence learning: comments on Stadler (1995) and Curran and Keele (1993), *J. Exp. Psychol.: Learn. Mem. Cogn.* 24 (1998) 234–248.
- [9] P.F. Dominey, M.A. Arbib, J.P. Joseph, A model of cortico-striatal plasticity for learning oculomotor associations and sequences, *J. Cogn. Neurosci.* 7 (3) (1995) 311–336.
- [10] P.F. Dominey, Complex sensory-motor sequence learning based on recurrent state – representation and reinforcement learning, *Biol. Cybern.* 73 (1995) 265–274.
- [11] J.L. Elman, Finding structure in time, *Cogn. Sci.* 14 (1990) 179–211.
- [12] A. Gillies, G. Arbutnot, Computational models of the basal ganglia, *Mov. Disord.* 15 (5) (2000) 762–770.
- [13] D. Joel, Y. Niv, E. Ruppin, Actor–critic models of the basal ganglia: new anatomical and computational perspectives, *Neural Networks* 15 (2002) 535–547.
- [14] T.L. Luks, H.C. Nusbaum, J. Levy, Hemispheric involvement in the perception of syntactic prosody is dynamically dependent on task demands, *Brain Lang.* 65 (2) (1998) 313–332.
- [15] L. Mavlov, Amusia due to rhythm agnosia in a musician with left hemisphere damage: a non-auditory supramodal defect, *Cortex* 16 (1980) 331–338.
- [16] Y. Morlec, G.L. Bailly, V. Aubergé, Generating prosodic attitudes in French: data, model, and evaluation, *Speech Commun.* 33 (2001) 357–371.
- [17] T. Nazzi, J. Bertoncini, J. Mehler, Language discrimination by newborns: towards an understanding of the role of rhythm, *J. Exp. Psychol.: Hum. Percept. Perform.* 24 (1998) 1–11.
- [18] B.A. Pearlmutter, Gradient calculation for dynamic recurrent neural networks: a survey, *IEEE Trans. Neural Networks* 6 (5) (1995) 1212–1228.
- [19] I. Peretz, Processing of local and global musical information by unilateral braindamaged patients, *Brain* 113 (1990) 1185–1205.
- [20] F. Ramus, J. Mehler, Language identification with suprasegmental cues: a study based on speech resynthesis, *J. Acoust. Soc. Am.* 105 (1) (1999) 512–521.
- [21] J.R. Saffran, R.N. Aslin, E.L. Newport, Statistical learning by 8-month-old infants, *Science* 274 (5294) (1996) 1926–1928.
- [22] R. Shi, J.L. Morgan, P. Allopenna, Phonological and acoustic bases for early grammatical category assignment: a cross-linguistic perspective, *J. Child Lang.* 25 (1998) 169–201.
- [23] R.E. Suri, W. Schultz, Learning of sequential movements by neural network model with dopamine-like reinforcement signal, *Exp. Brain Res.* 121 (1998) 350–354.