Learning to Perceive and Produce Rhythmic Patterns in an Artificial Neural Network

J. Devin McAuley[1]
Department of Computer Science
Indiana University
Bloomington, Indiana 47405

February 1, 1993

[1]The author is supported by ONR grant N00014-91-J1261. Thanks to Michael Gasser, Robert Port, Joseph Stampfli, Jonathan Mills, Paul Purdom, Sven Anderson, Gary McGraw and Catherine Rogers for their constructive comments and criticism during the development of this manuscript.

1 Introduction

The perception and production of rhythmic patterns is a complex temporal process that characterizes fundamental aspects of human activity, interaction, and intelligence. For example, walking, running, dancing, speaking, and playing a sport or musical instrument involve *both* perception and production of rhythm. In a basketball game, players must anticipate the actions of their teammates. One way to do this is to perceive a player's movements as a rhythmic pattern; but, this is only possible if the team members are in fact moving rhythmically. This example is suggested in Handel (1989):

> The rhythm of a game (of basketball) emerges from the rhythm of individuals, the rhythm among team members, and the rhythmic contrasts between opposing teams.

Human rhythm perception/production is *relatively* easy; i.e. once learned, walking, running, dancing, and speaking become automatic processes, requiring very little conscious attention. In contrast, machine rhythm perception/production is an unsolved problem which has for the most part been overlooked in the field of Artificial Intelligence (AI).

James Anderson (1992) makes an interesting observation about flexibility and accuracy in human learning of multiplication facts which applies to the problem of machine perception of rhythms. Parallel machines are extremely fast and precise; yet humans, who have sophisticated and massively parallel "wet-ware", are slow and inaccurate at multiplication. Why is this? The human brain as an *accurate* symbolic processor is a recent evolutionary development. It is built on top of very *flexible* perceptual mechanisms, refined by evolution for billions of years. I claim then that human and machine information processing emphasize very different mechanisms and that perception has a lot more to say about human intelligence than symbolic logic does. Consequently, more research emphasis in AI needs to be placed on perception.

The emphasis of this thesis proposal is on the development of a connectionist model which learns to perceive and produce musical rhythms. Music offers a very rich but manageable domain for studying rhythm. I anticipate that the general mechanisms developed in the thesis will also contribute to research in speech and locomotion.

2 The Problem of Variability

Variability is the main reason why rhythm perception is a difficult AI problem. To make this point more concrete, imagine that two people are playing a rhythm game roughly analogous to "Simon Says". The first player's task is to invent novel sequences of finger taps. The goal of the second player is to reconstruct this pattern as accurately as possible, by tapping. A colorful musical interpretation of this interactive perception-production task is the bluegrass piece "dueling banjos" by Lester Scrugs and Earl Flatt. This tune begins with a very simple theme which is played first by a guitarist and then replayed by a banjo player. As the piece progresses, players exchange increasingly complex variations on the theme. By the end of piece, both musicians have joined forces and the original theme can now be heard as part of their combined improvisation.

In both of these examples, there is substantial production to production variability. For even the most sophisticated musician, it is impossible to reproduce temporal intervals *exactly*. There is local variability in tap/tone onset times, and there is global variability in production rate, intensity level, and timbre (tapping surface, guitar versus banjo). In spite of this variability, human listeners easily classify different productions as the same rhythmic pattern. The main point is that rhythm perception and production is not a simple process of memorizing temporal intervals. Variation in production forces flexible perception.

Variability is also a major problem is speech recognition. There is substantial between-talker and within-talker variability in different productions of the same utterance. Productions vary in speaking rate, intensity level, intonation, timbre, and context. Background noise further confounds the problem.

A good example of the inability of traditional rule-based systems to handle variability is the HEARSAY II model for speech understanding, developed by Lesser et al. (1975). Although very successful for its general contributions to the field of AI, i.e. introducing the concept of a "blackboard" architecture and competing knowledge sources, it was not very successful as a speech recognizer. HEARSAY II achieved a high degree of accuracy, but only on a limited vocabulary finite state grammar constrained to a document retrieval task. Recognition was further restricted to the voice of one speaker in a relatively quiet environment.

Solving the variability problem requires introducing greater flexibility into a model than can be obtained with traditional rule-based approaches such as HEARSAY II.

3 Factors Affecting Rhythm Perception

Handel (1989) defines rhythm as the "sense of a regular, periodic sequence of subjectively stronger and weaker beats". The sequence of strong and weak beats constitutes rhythmic structure and is often called a *metric hierarchy*. The strength of a beat is determined by the number of beats it overlaps with in the hierarchy. A two beat meter hierarchy for an eight- note repeating pattern is shown in figure 1. The first element is the strongest beat because

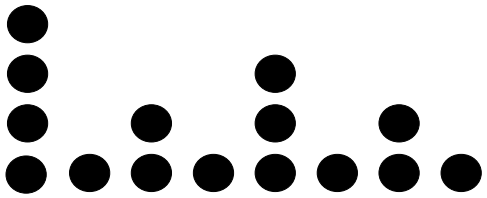

Figure 1: A Two Beat Metric Hierarchy

the hierarchy overlaps on four levels. The fifth element, beginning the second group of four, is the next strongest beat because the hierarchy overlaps on three levels.

1. Early research in rhythm perception (Bolton, 1894) explored human perception of sequences of physically identical elements separated by equal intervals. Without physical cues in the signal, subjects tend to group elements by two's, three's or four's. This phenomenon has been labelled *subjective rhythmization* and reveals that context plays an important role in rhythm perception. A model for rhythm perception should allow biases to affect the formation of percepts.

2. In a review of rhythm perception research, Fraisse (1978) reports that if the separation between two successive tones is less than 50 msec, then the two elements are not perceived as separate beats. If the separation between tones is greater than approximately $1.5 - 2.0$ seconds, then

the two events are heard as disjoint, i.e. not belonging to the same sequence. Thus, an apriori constraint on the model is that it needs to be able to encode frequencies between 0.5 Hertz and 20 Hertz.

3. In earlier work, Fraisse (1963) explored whether or not human subjects have a "spontaneous tempo". Subjects were asked to tap at their preferred rate. As might be expected, there are large rate differences across individuals, ranging from 1.1 Hertz to 5.0 Hertz. When subjects were asked to reproduce temporal intervals, accuracy was maximal for intervals around 600 msec (1.7 Hz). This is often used as a representative tapping rate.

4. Many researchers have explored the effects of variations in intensity, duration, pitch, and timbre on rhythm perception. Subtle changes in all of these factors have been shown to affect grouping. For example, if every second element of an isochronous sequence is accented by increasing its intensity, then the sequence is grouped by two's with the accented element beginning the group. Instead if every third element is increased in intensity, grouping is by three's. Changes in element duration have been shown to have similar effects to changes in intensity, except that it is the last element of the group that is perceived as accented. Changes in several factors at once have been shown to have interacting effects that either facilitate or inhibit group formation. This research is summarized in Handel (1989).

5. People tend to lock onto perceived rhythm by being able to predict future input in terms of the underlying beat structure. For example, the underlying beat of a musical pattern continues to be heard even when there are rests or the pattern stops altogether.

6. Bregman and Cambell (1971) observed that grouping in auditory perception obeys many of the principles of Gestalt organization. Principles such as similarity, continuity, and common fate apply to groups of elements that are not necessarily adjacent in time. For example, a sequence of alternating high and low tones can be perceived as two distinct sequence *streams* consisting of high tones only and low tones only. Presentation rate affects the formation and merging of streams.

To summarize, I have briefly discussed a number of perceptual factors in rhythm perception to provide some indication of its subtleness. The context of a local change can bring about global temporal reorganization of the entire sequence percept. There is a "lack of correspondence between the characteristics of the physical acoustic wave and the perceived rhythm and accentuation (Handel, 1989, p. 390)." Rhythm perception is an emergent phenomenon.

4 The Model

The goal of this research proposal is *not* to incorporate into a system all of the subtle factors involved in rhythm perception; i.e. in the proposed system, I am not interested in solving the problem of auditory streaming or modelling the *quantitative* results of Fraisse, Povel and Essens, and others. Instead, the aim is to explore the question of how a model might self-organize flexible perceptual representations for rhythm that would be also useful in a production task, with emphasis on *qualitative* comparison to experimental work. The model should suppress noise, generalize over presentation rate, but maintain the interaction of some of the subtle grouping effects such as intensity accentuation. With respect to production, the model should be able to lock onto a perceived rhythm by being able to predict/produce future input in terms of the underlying beat structure. Rhythms identified through perception should enable production by "activating" the corresponding identification unit.

The description of the model is in four subsections. First, I will provide a very brief overview of the connectionist paradigm. Next, I will discuss two important implementation issues: how to represent input and time. I argue that input representations should be tied closely to the stimulus environment and that temporal information should be coded directly, as is the case in the mammalian auditory system. In the third subsection, I will motivate and discuss the two biological building blocks used in this proposal: pacemaker neurons and self-organizing feature maps. Finally, I will describe the network architecture in detail. The emphasis of this discussion will be on the self-organization of dynamic representations for prototype rhythms.

4.1 Neurally-Inspired Models

In the last decade or so, there has been a growing trend in the AI community to focus attention on more flexible research paradigms such as connectionist models which emphasize "brain-style" subsymbolic computation. The connectionist paradigm is based on simplifying assumptions about the architecture of the brain. The fundamental processing unit is a "neuron-like" element which receives a coded input signal via afferent "synaptic" connections. Each unit computes a function of the total input and transmits this signal to other processing units. Learning algorithms adjust the behavior of a network of processing units by modifying their synaptic weights. For a comprehensive introduction to connectionist models see Rumelhart and McClelland (1986) and McClelland and Rumelhart (1986).

Neural Networks have been particular successful at solving problems in pattern recognition partially because of their ability to generalize. By repeated application of a learning algorithm, a network codes the statistical *regularities* of the training set in its synaptic weights. It generalizes by classifying novel input patterns based on these coded regularities. Distributing processing over many units introduces computational redundancy. Consequently, connectionist models tend to be fault-tolerant with respect to processing unit damage and impoverished input. Rules emerge as a by-product of subsymbolic interactions.

4.2 Representation

In an Introductory Artificial Intelligence textbook, Charniak and McDermott (1985) define a representation as "a stylized version of the world." In principle, a representational framework should reduce the quantity of stimulus information to a manageable size, but preserve its quality. The problem with this principle is that it is difficult, if not impossible, to decide what information can be thrown out.

Deciding on an input representation *is* the problem of perception. It is often sidestepped in AI by assuming preprocessed input. Preprocessors are most likely unspecified and the resulting input representations reflect arbitrary decisions. Orthogonal vector representations are a good example. Such arbitrary representations do no reflect properties and relationships between stimuli; i.e., the relationship between stimulus A and stimulus B does not

correspond to the relationship between the representation for stimulus A and the representation for stimulus B. Claims by such a model, especially about perception, are difficult to support. For this reason, I argue that input representation should not reflect arbitrary decisions about salience, but instead show similarity to the raw stimulus.

By definition, rhythm perception is a temporal process. To adequately recognize, represent, and generate rhythmic patterns, a model needs to be able to represent detailed temporal information about the stimulus. I argue that to do so, a system should directly code temporal information; i.e., time should be a parameter of the model.

This representation should not be a static "picture" of the stimulus, as there is a clear left-to-right asymmetry in the perception of temporal events. For example, frequency discrimination on a particular "target" component of a 10 tone pattern is more difficult when the pattern is played forwards than when the pattern is played backwards (Charles Watson, 1992). A static spatial representation for temporal events does not capture this asymmetry, as forwards and backwards do not have functional significance.

In most AI models which do consider the importance of *left-to-rightness* in temporal processing, the representation of time corresponds to an abstract sequence of events such as the alphabet. [2] Sequence elements precede and follow other elements, but the representation does not encode quantitative information about duration, necessary for perceiving rhythms.

On the other hand, the representation of temporal events in the mammalian auditory system is more detailed and suggests direct durational encoding. From the auditory nerve to the auditory cortex, a consistent set of cell-response types have been observed for temporal patterns of stimulation: "through" cells maintain firing during presentation of the stimulus, "on" cells respond to stimulus onset, "off" cells respond to stimulus offset, and "on-off" cells respond to both stimulus onset and offset (Pickles, 1988). In addition, there are cells which tend to fire whenever the stimulus frequency reaches a particular phase angle (Moore, 1982). [3]

A subgoal of the proposed research is to focus attention on the temporal information available to the auditory system–and biological systems in

[2] Two examples of connectionist architectures which are often associated with sequential representations are the **SRN** (simple recurrent network) of Elman (1990) and the **RAAM** model (recursive auto-associative memory) of Pollack (1988).

[3] The representation of temporal patterns is discussed in detail by Port (1990).

general–in designing a model for rhythm perception and production.

4.3 Building Blocks from Biology

4.3.1 Pacemaker Neurons

One of the simplifying assumptions of the connectionist paradigm is that an activation value can be used to model the firing rate of a neuron. Since rhythm perception is a temporal process, this simplification is not an advantage. In the proposed model, processing units maintain temporal information by exhibiting oscillatory behavior.

The hypothesis that populations of oscillatory neurons or "pacemakers" code temporal information has been studied extensively by many researchers, but only at a very low level. John (1967) describes a number of these animal experiments. The basic experimental procedures involve variants on the classical conditioning paradigm.

Entrainment: Repeated pairing of a constant tone (unconditioned stimulus) and a flickering light (conditioned stimulus) eventually evokes electrical activity at the frequency of the flicker, once the conditioned response is obtained. Thus showing that the flicker frequency is a direct code for triggering the conditioned response. Initially, entrainment of the conditioned stimulus occurs only in the peripheral regions, but as the animal becomes more sensitive to the conditioned stimulus, the conditioned frequency appears in more central cortical regions as well.

Assimilation: If presentation of the conditioned stimulus stops, neurons continue to fire at the conditioned frequency. If pulses of the conditioned stimulus are left out, such as deleting flashes from a flickering light, then neurons fill in the missing "beats" by firing when a deleted pulse would have occurred.

Generalization: The presentation of a stimulus at a slightly different frequency than the conditioned one triggers the conditioned response and evokes a firing pattern corresponding to the conditioned stimulus in central cortical regions.

These three properties exhibited by populations of "pacemaker" neurons are essential features of the rhythm perception process. A model for rhythm needs to be able encode a particular rhythm (entrainment), predict when the next beat will occur after the rhythm has stopped (assimilation), and be able

to recognize the same rhythm at varying rates (generalization).

A number of pacemaker models have been proposed. One such model (Perkel et al., 1964) was extended by Torras (1985) to handle entrainment, assimilation, and generalization. Torras's extended model is the foundation of the present work. The *integrate-and-fire* model consists of four unit equations: an input activation function, a spontaneous activation function, a total activation function, and a variable threshold equation. The input at time t is given by $i(t)$ and decays exponentially with a decay rate of τ_x. The total input is given by

$$x(t) = x(t - \Delta t)e^{-\tau_x \Delta t} + i(t).$$

In the absence of input, the spontaneous activity of the pacemaker is an exponentially increasing function with decay rate τ_s:

$$s(t) = s_l + (s_0 - s_l)e^{-\tau_s t}$$

The behavior of $s(t)$ is equivalent to a leaky integrator with constant input. Spontaneous activation is initially s_0 and approaches an asymptotic limit of s_l. Total activation is simply the sum of total input and spontaneous activity

$$y(t) = s(x) + x(t).$$

The threshold equation is defined by

$$h(t) = h_l + (h_0 - h_l)e^{-\tau_h t}.$$

The initial threshold value is h_0 and decays exponentially to an asymptotic limit of h_l. The decay rate is given by τ_h.

When the total activation function exceeds the variable threshold value, the unit discharges and all equations are reset to their initial values. In the absence of external input, total activation is equal to the spontaneous activation function. Figure 2 shows a graph of the spontaneous activation function and the variable threshold. When these two functions intersect, the unit spontaneously fires. The spontaneous firing period can be determined by setting the spontaneous activation function and the threshold equation equal and solving for t. If we assume that the spontaneous decay rate and the threshold decay rate are equal ($\tau_s = \tau_h = \tau$) and let b (the band size) equal the difference of the spontaneous limit and the threshold limit ($s_l - h_l$)

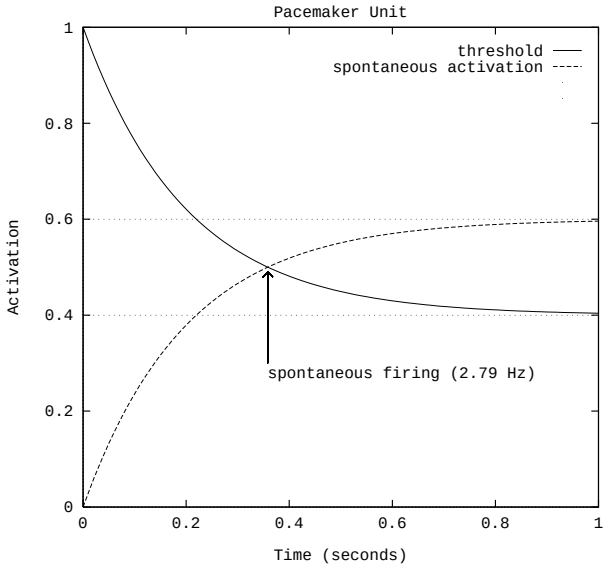

Figure 2: Pacemaker Unit. The exponentially increasing curve is the spontaneous activation function $s(t)$. Its asymptotic limit (s_l) is the upper dotted line (Activation = 0.6). The exponentially decreasing curve is the threshold equation $h(t)$. Its asymptotic limit (h_l) is the lower dotted line (Activation = 0.4). This unit spontaneously fires whenever the spontaneous activation function and the threshold equation intersect. The *time* coordinate of this intersection is the spontaneous firing period.

then the spontaneous firing period reduces to

$$t = \ln(\frac{b}{b+1})/-\tau.$$

In Figure 2, the band size is equal to 0.2. A graph of spontaneous firing rate versus band size is shown in Figure 3. The decay rate τ is 5.0.

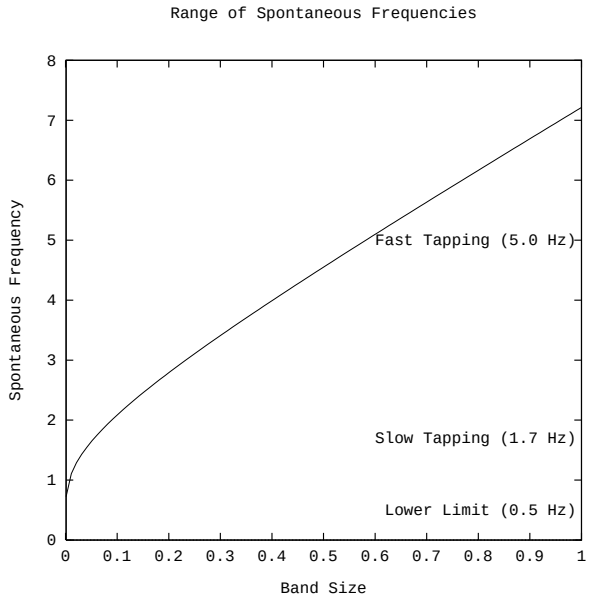

Figure 3: The spontaneous firing rate as a function of band size. The spontaneous tapping tempos determined by Fraisse are included to show the dynamic range of this pacemaker.

It's unclear whether real biological pacemakers can code the magnitude and range of rates associated with rhythm perception. That is, biological pacemakers often fire at rates that are orders of magnitude faster than musical rhythms. In an attempt to maintain consistency with the biology, Miall (1992) suggested that longer temporal durations might be coded by the beat frequency of a collection of coupled pacemakers.

Torras (1985) extends the Perkel model in three ways. Her first innovation is to model spontaneous discharge as a stochastic process. She does this by making the spontaneous asymptotic limit a Gaussian random variable with mean μ and standard deviation σ. Modest uncertainty in the spontaneous firing rate allows the pacemaker to handle local variation in input pulse arrival.

The second innovation is to allow the magnitude of a discharge to vary. This output function is given by

$$o(t) = o_{\max} - re^{h(t)-y(t)}$$

The output function is maximum whenever the input signal coincides with

spontaneous firing. In the absence of input, the output at the time of spontaneous firing is minimal. Consequently, in a network of units, only areas (collections of units) which correlate with the stimulus pattern will fire strongly.

The third innovation is to add an unsupervised learning rule. By stimulating a pacemaker at a particular frequency, it adjusts its spontaneous firing rate to match this frequency. The unit continues to fire at the stimulus frequency in the absence of stimulation. This enables a unit to predict when the next beat should arrive.

During learning, the arrival of an input pulse and the firing of a unit leave distinct *memory traces* in the state of the "neuron" that are only detectable above certain thresholds. If a unit fires when there is a detectable input trace, then the spontaneous firing rate is accelerated by increasing the slope of the spontaneous activation function so that the next time the unit discharges it coincides better the arrival of the input pulse. On the other hand, if an input pulse arrives soon after discharge– i.e. when there is a detectable discharge trace– the spontaneous firing rate is decelerated by decreasing the slope of the spontaneous activation function. In accelerative and decelerative learning the slope of the activation function is changed by increasing or decreasing the mean asymptotic limit of spontaneous activity by a constant value δ:

$$\mu = \mu \pm \delta.$$

The learning process is summarized in Figure 4. Requiring detectability thresholds fixes a "window" of entrainable frequencies centered around the spontaneous rate. The size of this window is determined by the magnitude of the input trace threshold and the discharge trace threshold. In general, increasing thresholds reduces window size. An input pulse which falls outside this window does not effect the behavior of the unit. Consequently, units can assimilate harmonic patterns as well as the fundamental. Since rhythms have harmonic structure, this property is an advantage.

4.3.2 Self-Organizing Feature Maps

In the nervous system, stimulus frequency is often a code for external events. For example, in the auditory pathway, there are fibers and cells which fire strongest for particular frequencies. Neighboring cells are selectively sensitive to neighboring frequencies. Although at the periphery tonotopic maps may

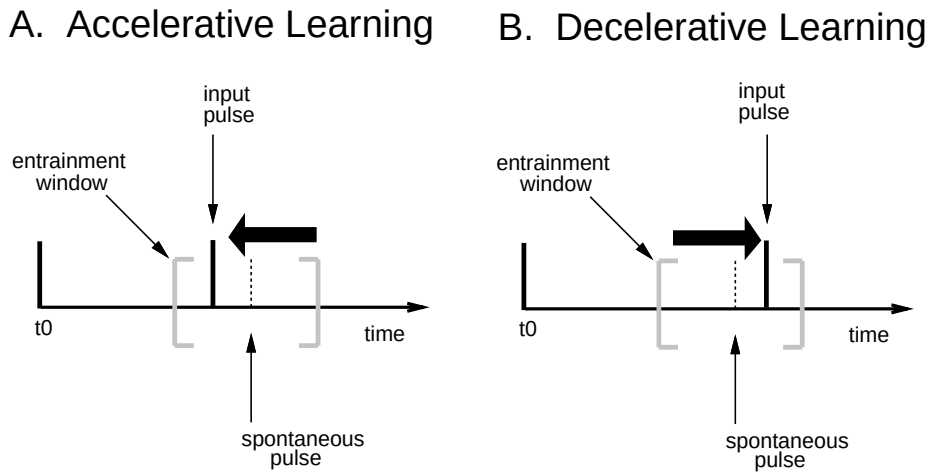

Figure 4: Windows of entrainment. In **A**, a pacemaker is stimulated with a pulse train that is faster than the spontaneous rate. If there is a detectable input trace at the time of spontaneous firing (the pulse falls within the accelerative side of the window), then the spontaneous rate is increasing by adding a constant to the spontaneous limit s_l. In **B**, a pacemaker is stimulated with a pulse train that is slower than the spontaneous rate. If there is a detectable firing trace when the input pulse arrives (the pulse falls within the decelerative side of the window), then the spontaneous rate is decreased by subtracting a constant from the spontaneous limit s_l.

be genetically predetermined, it is likely that in higher levels of the cortex these maps are self-organized. Malsburg (1973) presents a model for the self-organization of orientation-sensitive cells in the primary visual cortex. Using a simple learning rule, individual units uniquely code the orientation of "light bars" presented as input on a 19 unit retina. Clusters of units develop sensitivity to similar orientations.

Kohonen (1984) discusses a general algorithm for self-organizing feature maps such as those found in the primary visual cortex and the auditory pathway. The algorithm maps $n-$dimensional input vectors onto a two dimensional grid of output units. The weight vector associated with each output node is adjusted in such a way that topological neighbors develop sensitivity to physically similar inputs. Thus, unit responses reflect an ordering of the input space. The Kohonen algorithm is as follows. Weights to each output

unit are initialized to small random values. After an input is presented, the output unit with the largest activation value is determined either by hand (using a metric such as Euclidean distance) or by winner take all behavior associated with extensive lateral inhibition. The weight vector of the winning unit is adjusted so that it moves closer to the input vector. Weight vectors within a neighborhood of the winning unit are also adjusted so that they are closer to matching the input vector. Weight changes are inversely proportional to the distance from the winning unit; i.e., the weight vectors of distant neighbors are changed very little. Two conditions are required to obtain a convergent ordering of the input space: the neighborhood size and weight changes must decrease with time. Schyns (1991) implements these constraints with the following learning rule:

$$\Delta w_i = \begin{cases} \mathbf{x}(1 - o_w) f(o_i, o_w) & \text{for } i \text{ in } N_w \\ 0 & \text{for } i \text{ not in } N_w \end{cases}$$

In this equation, the weight vector for unit i is adjusted. The output of the winning unit is indicated by o_w; the set of units in the winner's neighborhood is N_w. Weight increment is scaled by the neighborhood function $f(o_i, o_w)$ which is a measure of the distance between the winning unit and neighboring unit i. As a unit develops sensitivity to an input pattern, its output becomes maximal and Δw_i approaches 0. As learning proceeds, the activation of the winning unit o_w saturates and the neighborhood size shrinks implicitly. Erwin et al. (1992) have shown that convergence rate can vary over several orders of magnitude depending on the shape of the neighborhood function, with convex neighborhood functions yielding the fastest convergence rates.

Three features of the Kohonen Architecture are important to mention in this discussion. **1.** Learning is a continuous process without distinct training and testing phases. Since we live in a dynamic environment, this is a distinct advantage. As a pattern becomes well-learned, weight increments approach zero and learning implicitly stops. Novel input patterns select winning units that have a relatively small magnitude, effectively triggering learning. **2.** Ordering of the input space assures that new input patterns will be coded by a unit that is near units already encoding a similar pattern. The problem of catastrophic forgetting in which a novel input pattern overwrites the memory of a well-learned pattern is thus avoided. **3.** If a set of units are lesioned from the network, the map will reorganize its representations.

4.4 Architecture and Learning

The proposed network consists of three functionally distinct layers as shown in figure 5. The bottom layer is a sensory-motor module consisting of a set

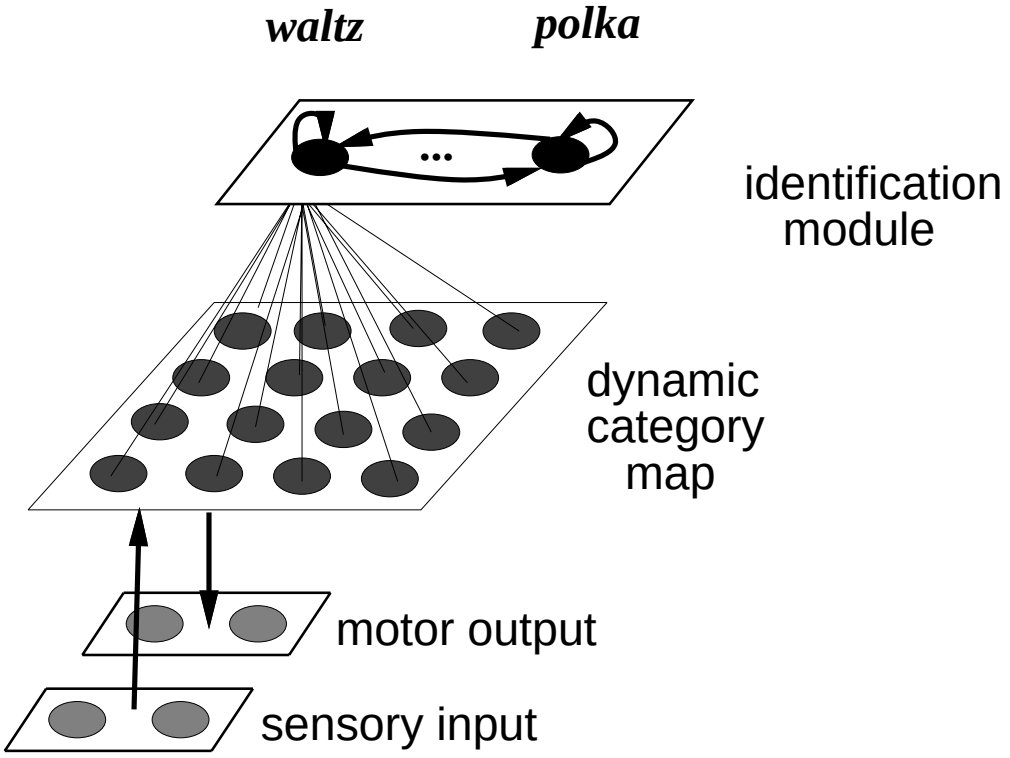

Figure 5: Network Architecture.

of acoustic input units and a set of motor output units. The middle layer is a *dynamic category* module combining the pacemaker units of Torras (1985) and the self-organizing feature maps of Kohonen (1984). The purpose of this module is to develop dynamic representations for prototype rhythms such as *waltz* or *polka*. This is accomplished by an algorithm which self-organizes an oscillator map reflecting the rhythmic structure of the stimulus environment. The top layer is an identification module. Using a supervised paradigm, "names" are attached to the prototype rhythms; e.g., this pattern of activation corresponds to a waltz. Activating a *name* unit excites feedback connections into the category module, and consequently triggers oscillatory

behavior corresponding to the encoded prototype. At this level, perception and production share representation.

4.4.1 Sensory-Motor Module

Todd (1992) has developed a set of analog filters which preprocess musical phrases into a pattern of snare-drum-like pulses. In his system, input is filtered by a hierarchy of harmonically spaced Gaussian filters. The frequency range of these filters coincides with tempos in music. Band pass filtering in this range converts music into a pattern of "beats"; intensity and duration are preserved but pitch changes are eliminated.

Input to the network will be an unstructured pattern of pulses generated by a program intended to simulate aspects of Todd's filter system. Using a program to generate pulses will make it easier to test the models performance on changes in rate and shifts in intensity accentuation. The program interface will allow the user to specify intensity changes to individual pattern components or changes to the overall pattern structure such as presentation rate. The program will allow multiple channels which correspond to the acoustic input units of the network. Outputs will be in the range $[0, 1]$ with magnitude as a rough intensity measure.

The motor portion of this module which might be used to do synthesis or drive a motor task is not discussed in this proposal.

4.4.2 Dynamic Category Module

The dynamic category module combines self-organizing feature maps and Torras pacemakers. It consists of an *n-by-n* grid of pacemakers that are self-organized with respect to firing rate using a Kohonen-like algorithm. Each pacemaker receives input from all of the acoustic units. All weights to the category module are initialized to equal values with the sum of the weights equal to one; e.g., with two acoustic inputs, all weights are initialized to 0.5. An eventual goal is to adjust these connections during self-organization with initial weights set to normalized random values.

The learning algorithm is roughly analogous to the Schyns (1991) modified **SOFM** algorithm. Spontaneous firing rates are initialized to a range of medium-rate musical tempos. Window sizes are initially set to a fairly wide range of oscillation periods so that self-organization develops representations

for a wide range of musical rhythms. Stimulation and learning are a continuous process. Since self-organization is in time, the question of *when* a winner is selected becomes as important as *how* a winner is selected.

A winner is potentially determined on each input pulse. If an input pulse triggers pacemaker discharge, then the "winner" is the pacemaker with the largest output amplitude above a threshold $o_{\min}$. This threshold is set to be equal or larger than the magnitude of spontaneous firing so that learning does not occur in the absence of input. The winning unit is accelerated using the update rule

$$\mu_w = \mu_w + \frac{1 - o(t)}{1 - o_{\min}}\delta.$$

With respect to Torras's pacemaker model, it is an innovation to consider learning steps variable. If δ is scaled by some form of the term

$$(1 - o(t))$$

then learning adjusts the spontaneous firing rate proportionate to how well the spontaneous frequency and input frequency correlate. If the spontaneous rate is substantially slower than the stimulation rate, $o(t)$ is relatively small and the learning jump is close to its maximal value δ. With perfect entrainment, the output magnitude at discharge is close to one and the scaling factor is zero. That is, learning essentially stops once spontaneous frequency and input frequency match.

If an input pulse does not trigger unit discharge, learning is decelerative. The winning unit is the unit that has fired most recently (time T within time $T_{\max}$). The decelerative update rule is analogous to the accelerative rule:

$$\mu_w = \mu_w - \frac{T}{T_{\max}}\delta.$$

To compute T, time can be measured either directly using a clock or implicitly using the memory trace of the last discharge.

The spontaneous rates of units within a neighborhood of the winning unit are also adjusted. Spontaneous rate change of neighboring units is inversely proportional to the unit's distance from the winning unit; i.e., the spontaneous rate of distant neighbors is changed very little. The accelerative update rule for unit i within a neighborhood of the winning unit is

$$\mu_i = \mu_i + \frac{1 - o_w(t)}{1 - o_{\min}} f(o_w(t), o_i(t))\delta.$$

As the winning unit becomes entrained to the stimulus pattern, the output of the winning unit $o_w(t)$ saturates and the self-organization of neighboring units becomes less efficient; neighborhood size shrinks implicitly.

Window size also shrinks during progressive entrainment. This is depicted in Figure 6. As a unit becomes more aligned to a stimulus pattern, its window of entrainable frequencies shrinks by increasing the thresholds for accelerative and decelerative learning. The rate at which the window shrinks is proportionate to the size of the learning jump. This is so that learning does not allow a unit to capture frequencies it could not have initially.

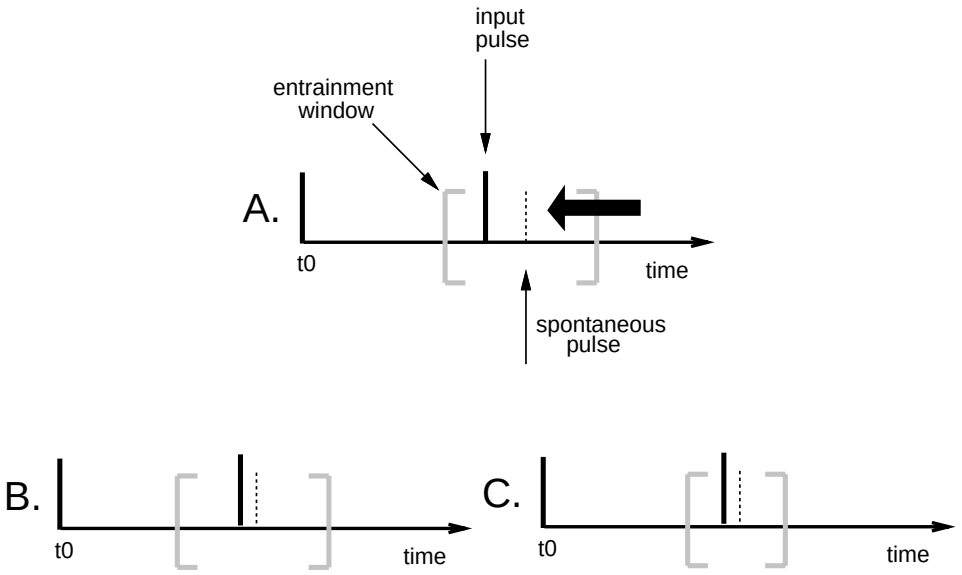

Figure 6: A unit is selected, using a winner-take-all procedure, to encode the frequency of a pulse train. **A** is prior to accelerative learning. **B** is after accelerative learning *without* shrinking the entrainment window size. Notice, that frequencies previously unencodable by the unit are now entrainable because the window has been shifted to the left without reducing its size. **C** is after accelerative learning, *with* the window shrinking proportionate to the size of the learning jump. The unit can no longer capture previously unencodable frequencies.

Figure 7 summarizes the self-organization process. It shows a cluster of pacemakers surrounding a unit entrained to frequency **A**. Unit size represents

the window size of the unit. Units with small windows are likely to discharge for only stimulation frequencies close to their spontaneous rate. Units with large window sizes are less selective. Shading shows how well a unit is entrained to the stimulation frequency. The center unit with black shading is perfectly entrained, except for stochastic fluctuations in spontaneous firing. The dark gray units fire at frequencies similar to to **A**. A second stimulation frequency **B** similar to **A** would be encoded by one of the dark gray units close to the **A** entrained unit. A third stimulation frequency **C** very different from **A** or **B** would be encoded by a unit farther away. Eventually, a topological ordering of oscillators is obtained.

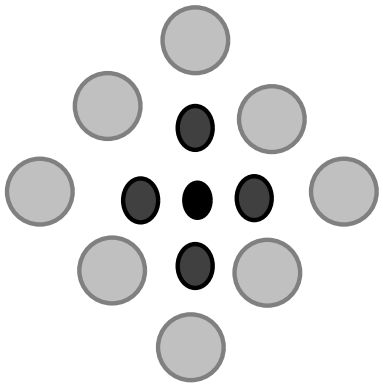

Figure 7: Self-organized response to one frequency.

There is a problem with the self-organization algorithm as it is stated. With respect to rhythm perception, a nice property of pacemaker units is that coupling pacemakers with harmonic frequencies reinforces oscillations. With respect to self-organization, this property is undesirable because a topological ordering of oscillators requires 1-1 entrainment. In the current pacemaker model, it is as likely that a unit will entrain to a harmonic pattern as it will entrain to the fundamental. Possible solutions to this problem include using a stochastic process for determining winners which allows multiple units to win, or adjusting how a winner is determined so that only units with potential 1-1 entrainment can win.

A related problem is the desirability that a rhythmic pattern only excite harmonics that are consistent with it; e.g., a three-beat rhythmic pattern should not excite units oscillating in a two-beat pattern. Connections need

to be learned between pacemakers so that unwanted harmonics are "turned off".

Self-organization of pacemakers is on a two-dimensional grid, but the ordering of stimulus information is only along one dimension (oscillation rate). The second dimension is unnecessary. I am maintaining both representational dimensions in anticipation of having the second dimension describe grouping and/or phase properties of the stimulus.

Continuous presentation of rhythmic patterns to the network develops a topological ordering of oscillators in which spatial patterns of oscillatory activity encode rhythm prototypes. When presentation of a rhythmic pattern stops, the associated spatial representation *persists* in memory until the firing amplitudes of the oscillators producing this spatial pattern decay. Activating a prototype rhythm requires external stimulation. This seems unnecessary since it is possible for people to imagine rhythms without any external stimulation. How this might be corrected in the model is discussed in the description of the identification module.

4.4.3 Identification Module

Units in the identification module are leaky integrators. Each unit integrates the series of oscillating spatial patterns representing a rhythm. Using a supervised paradigm such as the delta rule, *names* can be attached to rhythms encoded by the category module. A unit trained to "turn on" (have an activation of 1) during the presentation of a set of polka exemplars and "turn off" (have an activation of 0) during the presentation of a set of distractor patterns becomes a polka identifier. Lateral and recurrent on-center off-surround connections might be added to allow interaction between identification units.

Learning might also involve modulating feedback connections, which increases the oscillation magnitude of the pacemakers encoding the named rhythm. This would enable the model to hallucinate rhythmic patterns. For example, activating the waltz node would send an amplifying signal to all oscillators coding the prototype waltz. Such a strategy would also allow perceptual biases in map development such as turning on the waltz node so that even the slightest hint of a three-beat rhythm in a stimulus pattern is detected.

5 Simulations and Evaluation

In this section I will discuss a series of simulations aimed at exploring the performance of the proposed network in successive stages of its development. The first set of simulations explores the issue of self-organizing feature maps in time. The second set of simulations explores the network's ability to perceive and produce rhythmic patterns. Comparison with human performance is the suggested method for evaluating the model.

5.1 Map Formation

The first set of simulations tests the model's ability to self-organize oscillator maps and addresses a number of computational issues. The initial data set consists of pulse patterns with varying frequency. Each pulse pattern is a periodic binary sequence such as 10101010. The simplest task is to present a continuous stream of pulse patterns to the network and examine how the network self-organizes. Assuming that the problem of 1-1 entrainment has been solved, a number of issues will be interesting to explore. The most obvious question is how well the map reflects a spatial ordering of the stimulus space. What is the range of pulse frequencies that are entrainable and how does this depend on parameters such as window size and the initial distribution of spontaneous rates? Moreover, what properties are required for convergence? Second-order issues address the problem of flexibility. How well does the model reorganize after lesioning?

The second stage of this set of simulations is to test self-organized response to multi-input pulse patterns similar to musical rhythms. Allowing multiple inputs, connections between pacemakers, and connections from the sensory-module to the dynamic category module will effect convergence properties.

5.2 Perception and Production

The second set of simulations explores network performance in rhythm perception and production tasks. Input for all of these simulations are pseudo-musical patterns generated by a program simulation of Todd's filter system as described in section 4.4.1. For the purposes of this discussion, I want to focus attention on two types of rhythmic patterns: two-beat "polka" rhythms

and three-beat "waltz" rhythms. Imagine a data set constructed of 50 waltz exemplars, 50 polka exemplars, and 100 distractors which are rhythmically anomalous. Within-category exemplars vary in presentation rate, tone onsets, and intensity level. To test generalization, a network with two identification units is trained to label 25 randomly chosen waltz exemplars as waltzes, 25 randomly chosen polka exemplars as polkas, and not label 50 of distractor patterns. The network is then tested on the remaining exemplars. If the network successfully encodes prototype rhythms as spatial patterns, then it should be able to correctly label the novel waltzes and polkas while ignoring the distractors. That is, performance should generalize to modest rate changes and variations in intensity and tone onsets. A foreseen problem is that humans show a much larger range of rate invariance than is possible in this model.

To explore sensitivity to grouping effects such as intensity accentuation, the data set in the previous simulation is modified by adding intensity accents to 25 of the waltz exemplars so that their human perception shifts from a three-beat interpretation to a two-beat interpretation. This new exemplar is a *wolka*. Similarly, intensity accents are added to 25 of the polka exemplars so that their human perception shifts from a two-beat interpretation to a three-beat interpretation. This new exemplar is a *paltz*. The network trained in the previous simulation is presented with 25 waltz-wolka and polka-paltz pairs. A network that is successful at detecting these subtle accentuation effects will shift its identification from waltz to polka in the waltz-wolka case and from polka to waltz in the polka-paltz case. In a pilot study (McAuley, 1992), I demonstrated that a bank of oscillators can shift its dominant oscillation frequency from a two-beat to three-beat rhythm by only making a subtle change in the accentuation pattern of the input.

The network trained on the waltz-polka identification task can be used to examine performance in a production task. If the waltz unit is clamped then the feedback connections trained during the identification task will cause high amplitude oscillations in areas of the map corresponding to the prototype waltz pattern. This temporal pattern might then be used as input to a motor process.

6 Contributions

By exploring how a machine model might self-organize flexible perceptual representations for rhythms that would also be useful in a production task, this research can contribute to a better understanding of possible mechanisms underlying human perception and production of rhythms. The representation and processing of temporal information is not often addressed in Artificial Intelligence. Time is viewed as a problem, but for rhythm perception time defines the problem. This research can contribute to Artificial Intelligence by offering insight into how to represent and process temporal information. An innovation of this proposal is the description of an algorithm which combines pacemaker units and the Kohonen algorithm. Convergence of the Kohonen algorithm has only been proved in one dimension. This research can contribute to the theory of self-organizing systems by exploring the convergence properties of the proposed algorithm. By exploring important issues such as fault-tolerance and real-time information processing, this research can contribute to Computer Science.

A The Future of Computing

Since my thesis proposal could be classified under applied research, I found the following quote very supportive. In an effort to aptly define the future focus of Computer Science, the Computer Science and Telecommunications Board and the National Research Council (1992) have made recommendations to Universities regarding research.

> Academic Computer Science and Engineering should broaden its research horizons, embracing as legitimate and cogent not just research in core areas – but also research in problem domains that derive from nonroutine computer applications in other fields. The academic Computer Science and Engineering community should regard as scholarship any activity that results in significant new knowledge and demonstrable intellectual achievement, without regard for whether the activity is related to a particular application or whether it falls into the traditional categories of basic research, applied research, or development.

References

[1] James Anderson. Teaching the multiplication tables to a neural network: Flexibility vs. accuracy. Presented at the Computational Neuroscience Symposium, 1992.

[2] Sven Anderson, Robert Port, and J. Devin McAuley. Dynamic memory: A model for auditory pattern recognition. Unpublished manuscript, 1991.

[3] T. L. Bolton. Rhythm. *American Journal of Psychology*, 6(2):145–238, 1894.

[4] A. S. Bregman and J. Campbell. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, 89:244–249, 1971.

[5] Eugene Charniak and Drew McDermott. *Introduction to Artificial Intelligence*. Addison-Wesley Publishing Company, 1985.

[6] National Research Council. *Computing the Future: A Broader Agenda for Computer Science and Engineering*. National Academic Press, 1992.

[7] S. Dehaene, J.-P. Changeux, and J.-.P Nadal. Neural networks that learn temporal sequences by selection. In *Proceedings of the National Academy of Science*, volume 84, page 2727. National Academy of Science, May 1987.

[8] Jeffrey Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[9] Jeffrey L. Elman and James L. McClelland. Interactive processes in speech perception: The TRACE model. In James McClelland and David Rumelhart, editors, *Parallel Distributed Processing, Vol. 2*, pages 58–121. The MIT Press, Cambridge, MA, 1986.

[10] E. Erwin, K. Obermayer, and K. Schulten. Self-organizing maps: stationary states, metastability and convergence rate. *Biological Cybernetics*, 67:35–45, 1992.

[11] Paul Fraisse. *The psychology of time.* Lowe and Brydone, London, 1963.

[12] Paul Fraisse. Time and rhythm perception. In Edward C. Carterette and Morton P. Friedman, editors, *Handbook of Perception VIII: Perceptual Coding*, pages 203–254. Academic Press, New York, 1978.

[13] Michael Gasser. Towards a connectionist model of the perception and production of rhythmic patterns. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1989.

[14] Stephen Handel. *Listening: An introduction to the perception of auditory events.* Bradford Books/MIT Press, Cambridge, Mass., 1989.

[15] E. Roy John. *Mechanisms of Memory.* Academic Press, New York, N.Y., 1967.

[16] T. Kohonen. *Self-Organization and Associative Memory.* Springer-Verlag, New York, 1984. 2nd ed. 1988.

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

[18] V. R. Lesser, R. D. Fennel, L. D. Erman, and D. R. Reddy. Organization of the Hearsay-II speech understanding system. *International Conference on Acoustics, Speech, and Signal Processing*, 23:11–23, 1975.

[19] J. Devin McAuley. A model for the perception of temporal patterns. In *Proceedings of the International Joint Conference on Neural Networks*, volume 3, pages 798–803. IEEE, 1992.

[20] James L. McClelland and E. Rumelhart, editors. *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, volume 2. MIT Press, Cambridge, MA, 1986.

[21] Christopher Miall. The storage of time intervals using oscillating neurons. *Neural Computation*, 1(3):359–371, 1989.

[22] Brian C. J. Moore. *An Introduction to Psychology of Hearing.* Harcourt Brace Jovanovich, third edition, 1989.

[23] Donald H. Perkel, Joseph H. Schulman, Theodore H. Bullock, George P. Moore, and Jose P. Segundo. Pacemaker neurons: Effects of regularly spaced synaptic input. *Science*, 145:61–63, 1964.

[24] James O. Pickles. *An Introduction to the Physiology of Hearing.* Academic Press, San Diego, CA, 1988.

[25] Jordan Pollack. Recursive auto-associative memory: Devising compositional distributed representations. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*, pages 33–39, Hillsdale, NJ, 1988. Lawrence Erlbaum Associates.

[26] Robert F. Port. Representation and recognition of temporal patterns. *Connection Science*, 2:151–176, 1990.

[27] Dirk-Jan Povel and Peter Essens. Perception of temporal patterns. *Music Perception*, 2, 1985.

[28] David Rumelhart and James McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of cognition*, volume 1. MIT Press, Cambridge, Massachusetts, 1986.

[29] Philippe G. Schyns. A modular neural network model of concept acquisition. *Cognitive Science*, 15:461–508, 1991.

[30] Neil P. McAngus Todd. Recovery of rhythmic structure for expressive signals. Presented at the 124th Meeting of the Acoustical Society of America, 1992.

[31] Carme Torras. *Temporal-Pattern Learning in Neural Models.* Springer Verlag, Berlin, 1985.

[32] Carme Torras. Neural network model with rhythm-assimilation capacity. *IEEE Transactions on Systems, Man, and Cybernetics*, 16:680–693, 1986.

[33] Alan Turing. Computing machinery and intelligence. In Edward A. Feigenbaum and Julian Feldman, editors, *Computers and Thought*, chapter 1, pages 11–38. McGraw-Hill, 1963.

[34] Charles Watson, 1992. (personal communication) Indiana University.