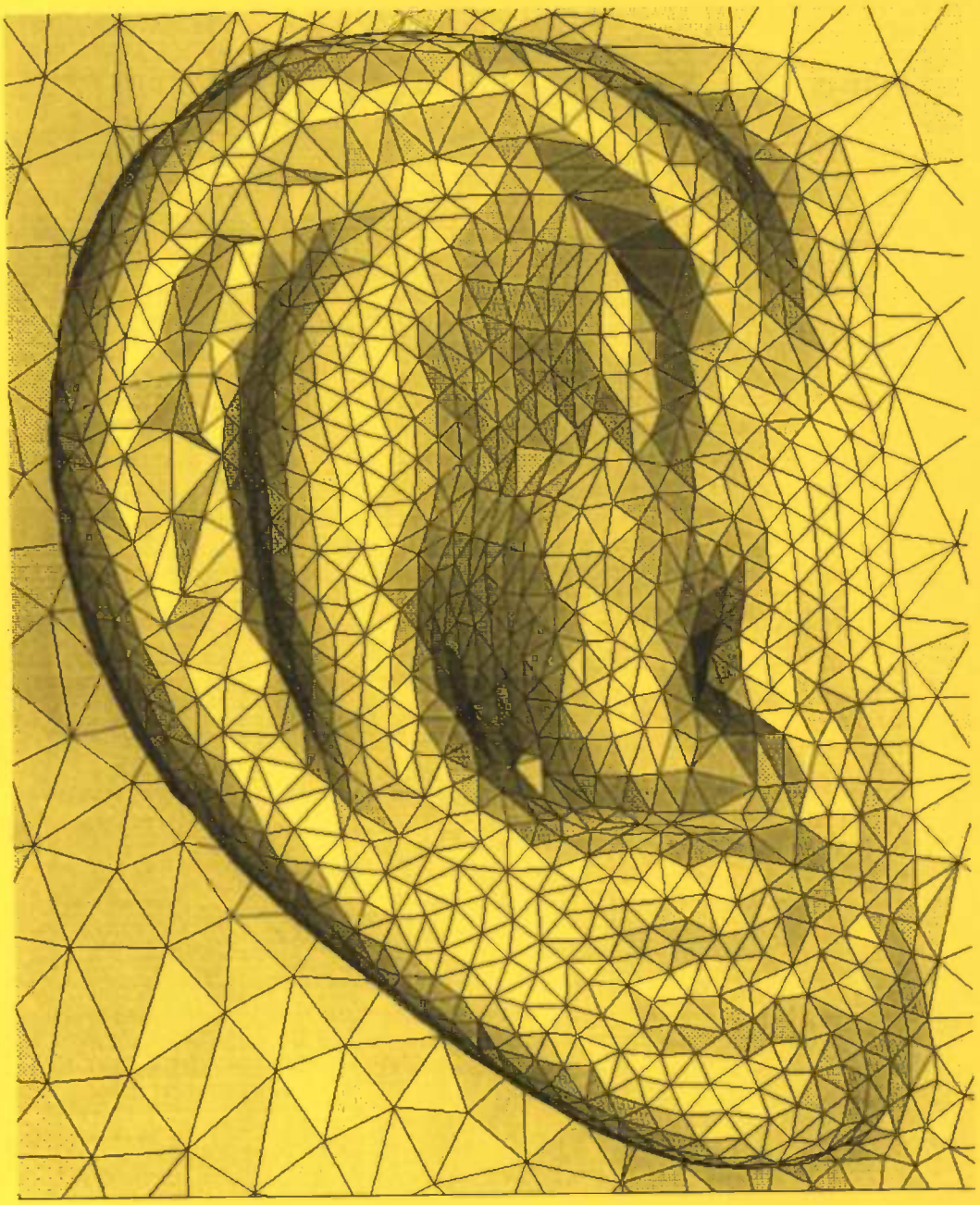
Timbre segregation and recognition in a model of the auditory system based on associative relaxation oscillator networks



Chris Honselaar (s0893358)
Fabian Gort (s0997935)

Rijks*universiteit* Groningen,
March 2002

Supervisors:
Dr. Ir. J.A.G. Nijhuis
Prof. Dr. P.G.M. Luiten

D 742

Rijksuniversiteit Groningen
Bibliotheek Biologisch Centrum
Kerklaan 30 — Postbus 14
9750 AA HAREN

BIBLIOTHEEK RU GRONINGEN
2228 2475

bij D 743

Erratum

bij "Timbre segregation and recognition in a model of the auditory system based on associative relaxation oscillator networks"

Bij de druk van dit verslag bleek helaas een belangrijk lettertype te ontbreken, waardoor de vergelijkingen onjuist zijn weergegeven. Hieronder volgen de correcties.

Pagina 8:

$$\frac{dx_i}{dt} = -\frac{x_i}{\tau_x} + G_x\left[T_{xx}\frac{x_i}{\bar{x}} - T_{xy}F\left(\frac{y_i}{\bar{y}}\right) + S_i^{exc} + S_i^{inh} + I_i - H_i\right]$$

$$\frac{dy_i}{d_t} = -\frac{y_i}{\tau_y} + G_y\left(T_{yx}\frac{x_i}{\bar{x}} - T_{yy}\frac{y_i}{\bar{y}}\right)$$

$$\frac{dH_i}{dt} = \alpha x_i - \beta H_i$$

(...)

τ_x and τ_y are time-constants.

Pagina 13:

$$\frac{dx}{dt} = f(x) - y + I + \rho$$

$$\frac{dy}{dt} = \varepsilon(g(x) - y)$$

where $f(x) = 3x - x^3 + 2$ and $g(x) = \alpha(1 + \tanh(x/\beta))$.
I represents the sum of lateral and external input to the oscillator, whereas α and β are parameters that determine the oscillation envelope and ε is necessary to achieve oscillatory state. Some pseudorandom noise was added by ρ for facilitating desynchronisation.

(...)

$$w_{ij} = \frac{1}{aN}\sum_{\nu}(\xi_i^{\nu} - a)(\xi_j^{\nu} - a)$$

(...)

ξ_i^{ν} is the set of patterns to be stored in the network. Total input for each oscillator is given by:

$$I_i = E_i + \sum w_{ij}\sigma(x_j)$$

(...)

$$\frac{dw_{ij}}{dt} = \gamma\sigma(x[t,i])(-w + \varphi\sigma(x[t,j]))$$

(...)

Rijksuniversiteit Groningen
Bibliotheek Biologisch Centrum
Kerklaan 30 — Postbus 14
9750 AA HAREN

$$I_i = E_i + \sum \frac{w_{ij} - w_{inh}}{a_2 N} \sigma(x_j)$$

Pagina 14:

$\sigma(x)$ is a sigmoid function: $1/(1 + \exp(-K*(x - \theta)))$
K is a parameter to adjust the steepness of the function, and θ is a threshold parameter.

(...)

Parameters used, except where mentioned:
common parameters: $\varepsilon = 0.02$, $\alpha = 6.0$, $\beta = 0.1$, $K = 50$
supervised model: $\theta = -0.5$, $a = 0.38$
unsupervised model: $a_2 = 1.3$, $\varphi = 0.8$, $\theta = 1.1$, $\gamma = 0.005$, $w_{inh} = 0.65$

Pagina 15, onderaan:

$$A(t,\tau) = \sum_{k=0}^{K=-1} r(t-k)r(t-k-\tau)$$

Pagina 16:

$$x_i = \sum_{j=i-d}^{j=i+d} q_j (1.4 - \frac{|j-i|}{2.72})$$

Pagina 18:

Figure 2. Pattern (0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0) is presented to the network and the original pattern is retrieved. Here, as in figures 8 and 9, $\theta = 2.02$.

D 743

Contents:

1. Introduction 4
1.1 The purpose of this investigation 4
1.2 Timbre segregation 5
1.3 Timbre recognition by associative memory 6
1.4 The peripheral auditory system 11
1.5 The central auditory system 12

2. Methods 13
2.1 Associative networks 13
2.2 Audio patterns 14
2.3 Inner hair cells 15
2.4 Segregation, autocorrelation and normalization 15

3. Results 17
3.1 Terman-Wang oscillator network (supervised) 17
3.2 Terman-Wang oscillator network (unsupervised) 22
3.3 Segregated patterns 26
3.4 Audio-pattern classification 28

4. Discussion 31
4.1 Average hamming distances 31
4.2 Sound recognition 31
4.3 Stability of oscillator dynamics 32
4.4 Pattern segmentation versus rebinding 32
4.5 Biological plausibility 33

5. List of references 36

Abstract:

For an animal it is important to be able to recognize meaningful sounds in the auditory scene. Sound recognition is largely based on timbre, a feature uniquely identifying a sound source irrelevant of its pitch. Therefore, the current research presents a model for the auditory system, which performs timbre processing, on the basis of periodic data. A periodicity-segregating network extracts the periodicities of experimental sounds and performs scale normalization in order to achieve pitch independent static audio patterns. After first exploring the capacities for pattern completion and segmentation of a Hebbian learning associative network of relaxation oscillators, it is modified for classification of audio patterns. The associative network appears to successfully segment and complete small patterns, but that does not apply to larger overlapping patterns. In three tests the neural network is trained with two instruments and then classification is performed. Results reveal a classification biased in the direction of one of the instruments. In addition, analysis of the segregated patterns suggests that periodicity alone cannot account for timbre recognition although it might provide some cues.

Abbreviations

AC Auditory Cortex · *AS* Auditory System · *AHD* Average Hamming Distance · *BM* Basilar Membrane · *BMF* Best Modulation Frequency · *HD* Hamming Distance · *IC* Inferior Colliculus · *IHC* Inner Hair Cell · *OHC Outer Hair Cell*

1. Introduction

1.1 The purpose of this investigation

It is important for an animal to be able to detect and recognize meaningful sounds in the auditory scene, like that of an approaching predator, or a congener in distress. There is, however, a continuous background noise from wind, rain, rustling leafs and so on. Furthermore, simultaneous sounds may interfere with one another as well. Thus, before an animal is able to recognize an independent sound it must first be segregated from the remaining audio scene. The auditory system (AS), which is part of the nervous system, is a specialized neural network that processes many aspects of sound and establishes effective segregation and identification of sound sources. In order to perform auditory scene analysis, the AS might make use of different sound features, among which are azimuth (horizontal location), loudness, pitch and timbre (sound colour).

As soon as a sound is segregated it is ready for recognition. An animal that has to make a judgement on the identity of a sound normally requires timbre information. The (rather poor) ANSI definition of timbre is: "... that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar". Psycho-acoustic investigations report three important dimensions that constitute timbre, which are attack, spectral shape and spectral flux (Krumhansl, 1989). Attack refers to the synchronicity and velocity of onset of the spectral components of a sound. Spectral shape describes the presence of frequencies and their relative amplitudes, whereas spectral flux represents the change of the spectral shape through time. In addition, context will influence timbre perception. With context cues are meant that influence the perception. Imagine a series of notes is presented to a listener, ascending to top and then descending again. Furthermore, all notes have an organ timbre, except for the top note, which has an oboe timbre. Under these circumstances the oboe tone will be perceived as an organ, just as the rest. Cues from other modalities, such as vision, may also influence timbre perception.

There are some practical problems concerning sound recognition, as the AS never receives the exact same input pattern twice (that is, new input varies in comparison to a previous stored pattern). In addition, it needs to handle degraded and distorted sound patterns as well. Wang et al. (1990) developed a biological plausible associative network model inspired by the animal olfactory system and based on plastic connections between oscillatory neurons, which solved these problems. In contrast with the majority of artificial networks, the oscillator network is able to express syntactical binding between sets of recognized features, by correlating their representation in the time domain. Instead of performing an 'all-or-nothing' classification of static patterns, it binds pattern elements by a mechanism of temporal correlation of oscillator activity, using familiar synaptic plasticity as proposed by Hebbs.

In this study a neural model of the AS is presented, which incorporates sound segregation and recognition, both on basis of spectral constancies in timbre. The animal brain requires processing of azimuth, loudness and pitch as well for segregation of sound patterns. These features are however not considered in the present model for reasons of simplicity. Timbre segregation is performed through a number of stages in a pre-processing network, whereas recognition is the task of an associative relaxation oscillator network. The latter is an adapted version of the above-mentioned Terman-Wang network, and implements an unsupervised Hebbian learning rule. During the first part of the experiment the capabilities of this oscillator network model are explored with small input patterns. In the second part the output of the auditory pre-processing system is presented to a larger oscillator network. Data of the segregation and oscillation network are analyzed independently, in order to assess the capacities of both. In conclusion, the biological plausibility of the proposed AS model is discussed. Before introducing the model, it is explained how segregation and recognition through timbre can be obtained.

1.2 Timbre segregation

A distinct sound is always perceived at a certain pitch. Spectral analysis of the waveform reveals that the sound wave repeats itself continuously and it appears that the wave's repetition frequency correlates with the perceived pitch. The frequency at which this 'overall wave' repeats is better known as the fundamental frequency (denoted as: F_0). Complex sounds, which are composed of layers of frequencies, might even lack a component at the fundamental frequency, while the pitch perception remains unchanged. Frequencies on top of F_0 are named partials and usually a number of partials are harmonic to F_0 (that is, their frequencies are integer multiples of F_0 (figure 1)). Although natural sounds always have some inharmonic components, there may be enough periodicity for the AS to associate them to a specific source.

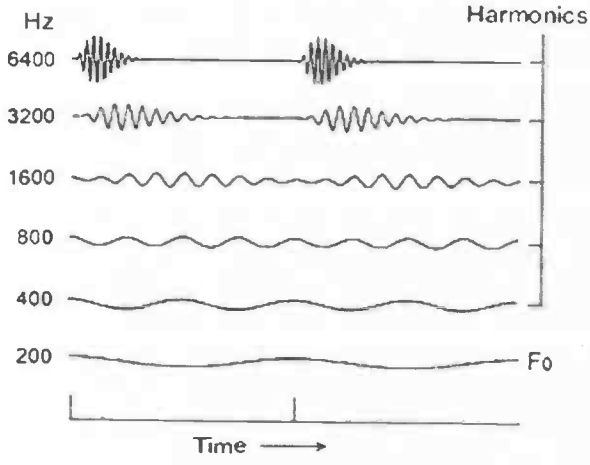


Figure 1. 2 Periods of a complex tone, consisting of an F_0 and 5 harmonic frequencies.

As mentioned, complex segregation of sounds requires more than timbre. This study merely focuses on separation of periodic and non-periodic information. A way to do this is by means of delay loops, which were proposed by Cariani (2001). A simple delay loop can be built from a signal delaying neural circuit, which loops the input. Neurons within the loop are coincidence detectors: they only generate an output when they simultaneously receive a spike from the input and from the loop self. When, for example, a spike pattern with a period of 15 ms is fed into a delay loop with a delay period of also 15 ms, then subsequent periods will coincidence within the loop. Each time two spikes coincide at the beginning of the loop it results in an amplified signal, which re-enters the loop. As a consequence, a periodic signal will build up in the delay loop. In contrast, re-entering spikes at the end of the loop, that do not encounter an input spike, will decay. This causes non-periodic information to be filtered out. Note how the length of a delay loop amplifying a periodic signal is inversely proportional to its fundamental frequency.

Autocorrelation can then be used in order to analyse periodicities from the delay loop output. Neural autocorrelation circuits utilize coincidence detection as well, but in a somewhat different way (Licklider, 1951). As can be seen in figure 2, these circuits consist of an input neuron that transmits signals to an instant line and a delay line. A spike from a temporal pattern will reach all the neurons in the instant line simultaneously, whereas it travels through the other line with a delay. As a consequence, during its trip along the delay line, it will encounter spikes in the instant line, which in the original pattern came after the concerned spike. Spike encounters are detected by a row of coincidence detectors, and so each coincidence neuron detects a unique spike interval. Feeding the output from a delay loop into an autocorrelator results in an interval distribution, which in turn corresponds to the periodicities present in the pattern. Phase information is lost in the process. However, this is in agreement with the findings that the relative phases of components have little effect on timbre recognition (Helmholtz, 1954).

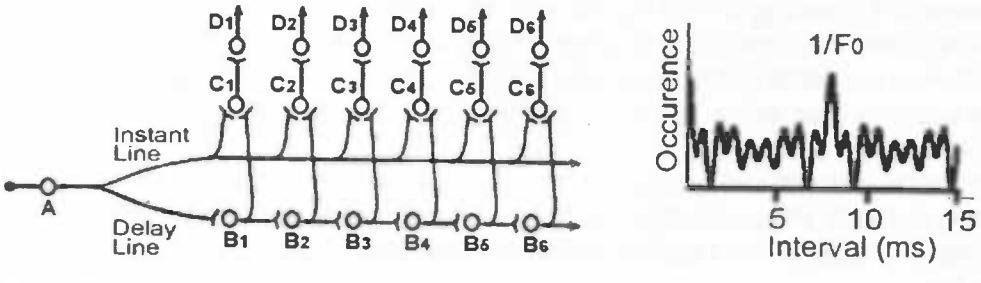


Figure 2. Left: Neuron A transmits its input to instant line and delay line. Neurons C1-C6 are coincidence detectors. Right: An autocorrelogram, which show an interval distribution. sound.

As sound recognition is at least partly pitch-independent, identification of timbre needs to deal with pitch-dependent variances in the periodicity distribution. This suggests that somewhere in the AS a normalization of timbre occurs. In this report, a method of timbre-normalization is proposed which effectively uses the spatio-temporal characteristics of the output of the delay

loop circuit to automatically generate a pitch-invariant representation of the periodicity distribution. Two tones at different pitches but with the same timbre will reveal the same spike pattern after this transposition.

1.3 Timbre recognition by associative memory

It is not yet known whether the animal brain stores auditory patterns within place-codes or whether temporal structure is preserved within synaptic delays. However, there is evidence that in the primary auditory cortex of Mongolian gerbils periodicities of 50-3000 Hz are coded by a place-code (Schulze and Langner, 1997).

As mentioned earlier, audio patterns processed by the AS never fully resemble previously stored patterns. Generalization is therefore a fundamental requirement for audio recognition. The classical Hopfield associative model solves these problems by storing memory traces as weighted connections between basic neural units, whereas static activity patterns form the short-term memory used for representation of the network state, such as memory retrieval. In dynamic systems theory, this model can be identified as an attractor neural network, since the short-term memory assumes dynamic attractor states (in the form of stationary patterns).
The main drawback of this kind of memory is the fact that it stores and retrieves patterns as indivisible entities, where all features of the input are lumped together. This hampers effective generalization especially for situations where multiple combinations of constant elements are likely to occur. Here, Hopfield networks and their variants have to allocate storage space for every element combination.
To achieve generalization by recognition of element combinations, a mechanism for syntactically binding these elements must be formulated. A well-known solution is to add neurons at a higher processing level to encode these combinations. However, this solution in turn leads to more problems ('grandmother-cell' dilemma).
Temporal correlations have been proposed to account for binding of common features. According to this hypothesis, temporal structure in the neural activity can encode symbolic linking between elements. Gray et al. (1989) found evidence supporting this idea, when they discovered neural oscillations in the cat visual cortex which showed synchronization in response to coherent, related stimuli.
Temporally structured activation patterns can be used as an alternative to static patterns in associative memory. For the specific case of oscillations, the attractor states of the Hopfield network have their counterpart in attractor limit cycles.
Oscillator networks constitute established neural processing models, and a number of categories are recognized, according to both characteristics of the connection and the oscillator dynamics themselves.

Phase-model oscillator networks are essentially abstractions of timing behavior observed in spiking neurons. They are not intended to capture actual synaptic dynamics, instead the oscillators directly influence the timing of connected oscillators.

Pulse-coupled (oscillator) networks approximate integrate-and-fire activity as found in many neurons. It can be shown that under certain circumstances the behavior of pulse-coupled networks can be mapped directly to that of phase-models, and even to the more detailed and biologically inspired Hodgkin-Huxley neurons (Izhikevich, 1999).
The oscillator model proposed by Wang (1990) for his associative oscillator network involves oscillator dynamics operating on multiple timescales: a resting state is alternated by an oscillatory 'bursting' mode. This phenomenon can be acquired by connecting inhibitory and excitatory neurons (or groups of neurons) in a feedback loop. For simulation purposes, Wang used the following set of equations:

$$\frac{dx_i}{dt} \quad \frac{x_i}{\Omega} \quad G_x \quad T_{xx} \frac{x_i}{\bar{x}} \quad T_{xy} F \quad \frac{y_i}{\bar{y}} \quad S_i^{exc} \quad S_i^{inh} \quad I_i \quad H_i \quad \cdots$$

$$\frac{dy_i}{d_t} \quad \frac{y_i}{\Omega} \quad G_y \quad T_{yx} \frac{x_i}{\bar{x}} \quad T_{yy} \frac{y_i}{\bar{y}}$$

$$\frac{dH_i}{dt} \quad \Delta x_i \quad EH_i$$

G en F are a sigmoidal gain function and a cubic function, respectively. x_i describes the average activity of the excitatory neurons, and y_i does the same for the inhibitory neurons, whereas $\bar{x}$ and $\bar{y}$ are parameters controlling the average values of x and y. The input from connected oscillators, modified by a weight value, is given by S^{exc}, for excitatory connections, and S^{inh}, for inhibitory connections. Ω and Ω are time-constants. Connection strengths between the inhibitory and excitatory neurons are given by T_{yx}, T_{xy}, T_{xx} and T_{yy}.

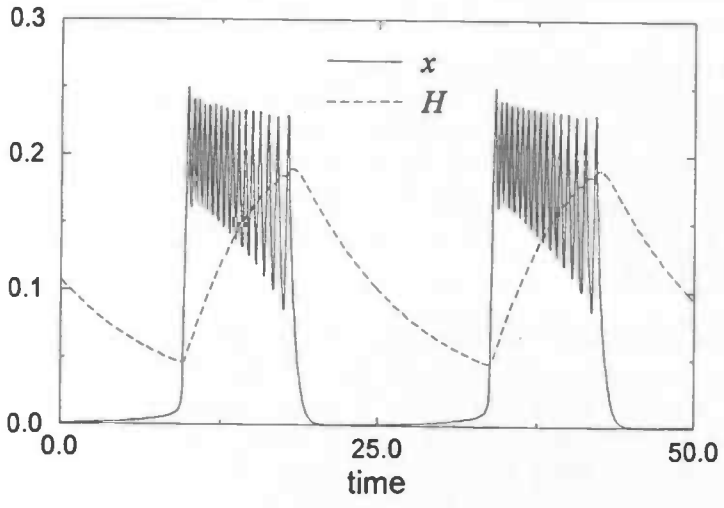


Figure 3. Time plot of excitatory neuron activity x, with H superimposed.

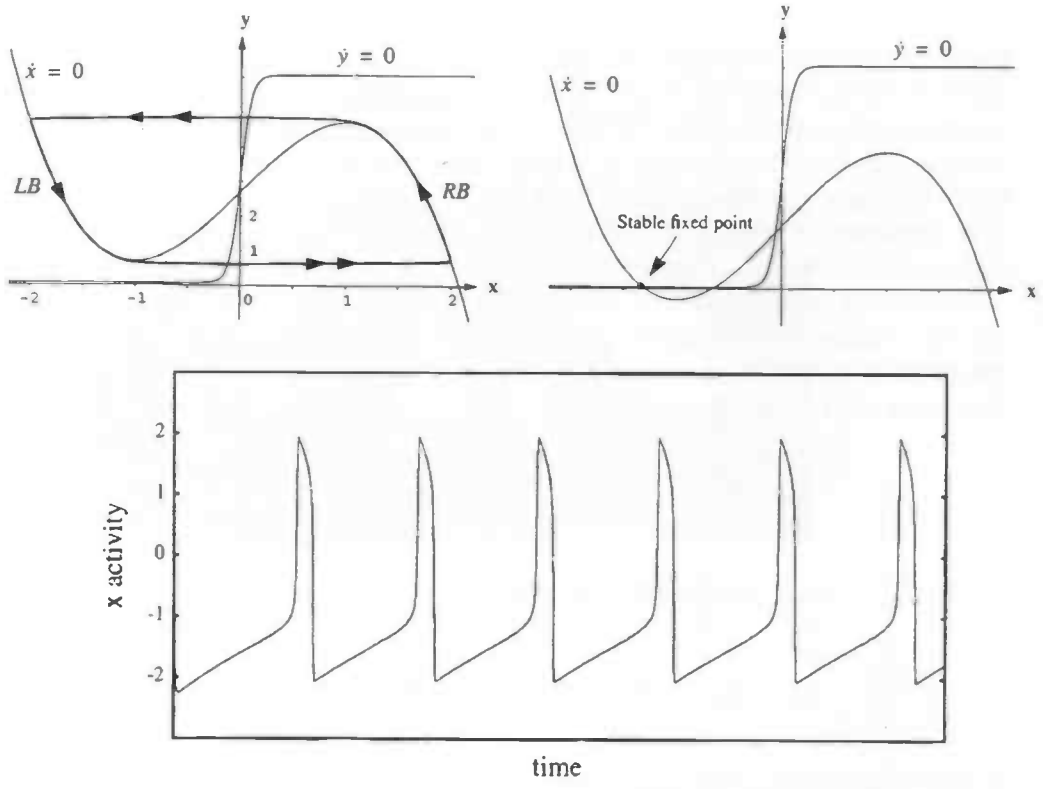


Figure 4. Dynamical behaviour of Wang-Terman oscillators. above: 2 phase diagrams displaying attractor limits in active state (left) and resting state (right). Below: typical time plot of oscillations.

These equations constitute a dynamical system with multiple asymptotic solutions. Stability of these solutions is influenced by H. Within a certain range of H bi-stability occurs for the attractors of the system x,y, causing a burst of oscillations. Figure 3 illustrates the dynamics for a typical set of parameters. The connection scheme proposed by Wang involves oscillators transferring a measure of their activity to the coupled oscillators, modified by a weight value. When mutual excitation between a pair of oscillators are considered, bursting in one oscillator results in a facilitation of the transition to bi-stability in the other oscillator, ultimately forcing them to synchronize their burst phases. An inhibitory connection on the other hand, will tend to cause desynchronisation between the coupled oscillators.

One of the most prevalent attempts to capture the laws governing the conductance-based membrane potential of nerve cells is the Hodgkin-Huxley model. For a certain set of parameters this model reveals oscillation on two timescales, switching between a resting and a spiking state. Generally, this is known as a relaxation oscillation. First discovered by van der Pol in experiments with triode circuits, it has been the basis for a number of models, among which are the FitzHugh-Nagumo equations and the Terman-Wang oscillator. The first is a simplification of the Hodgkin-Huxley equations, whereas the Terman-Wang oscillator was developed as a means of encapsulating this behavior, using easy-to-analyse mathematical equations

with a high degree of flexibility. The equations will be discussed in chapter 2.1.
The Wang-Terman oscillator consists of two dynamic components, which interact to form two basic behaviors: when the oscillator is unstimulated, the nullclines of these components will intersect in a stable point, keeping the oscillator in a rest state. If sufficient stimulation is present, the cubic nullcline is lifted, removing the stable intersection. Now, the system starts to cycle between the two legs of the cubic, where the left leg is the rest state and the right leg corresponds to an activity spike. The diagrams in figure 4 illustrate this.

A number of results have been proven about Wang-Terman oscillator networks. Importantly, it guarantees rapid synchronization in response to excitatory coupling, as well as stable desynchronized states. Networks of these oscillators have been applied in many practical experiments, such as auditory and image segmentation.

In this research, the capabilities of the Wang-Terman oscillator in an associative network are assessed. To determine its applicability as an associative memory, its performance is compared to that of the associative bursting oscillator network mentioned before. Lourenco et al. (2000) showed that this bursting oscillator network can be extended to self-organize in response to input. To this end, they proposed a learning rule encompassing elements of Hebbian synaptic plasticity. In this report the capacity for autonomous learning in Wang-Terman networks is researched by applying a standard Hebbian learning rule. Hebbian plasticity of synaptic connections between neurons was found in the auditory cortex (Ahissar et al., 1998)
In particular for biological sciences the associative oscillation network is rather interesting, since oscillatory activity of neural populations is a commonly observed phenomenon in the cortex. Oscillatory activity was also established in the auditory cortex (Sukov et al., 2001).

Returning to the problem of timbre recognition, it is important to note that the classes of associative networks discussed so far all perform generalization over Hamming Distance. That is, differences between patterns are evaluated in terms of different elements at corresponding positions. This entails that any variance in position, scale or other dimensions of the patterns will tend to lead to faulty recognition. Thus, in order to perform robust timbre recognition, one has to ensure that variances like these are eliminated in the preprocessing stage. As mentioned, the present model adopts scale-normalization in order to obtain size-invariant patterns.
Time is another factor complicating recognition. The identification of temporal patterns requires some method for detecting sequences or higher-order time structures. When considering timbre features, an important time structure is to be found in the periodicities of the auditory signal, as discussed earlier.
In the preprocessing stage presented here, autocorrelation is performed on auditory data, which extracts the periodicity features, and effectively discards random temporal patterns. Additionally, temporal integration over sufficiently large timeframes yields patterns that are static enough for recognition processes.

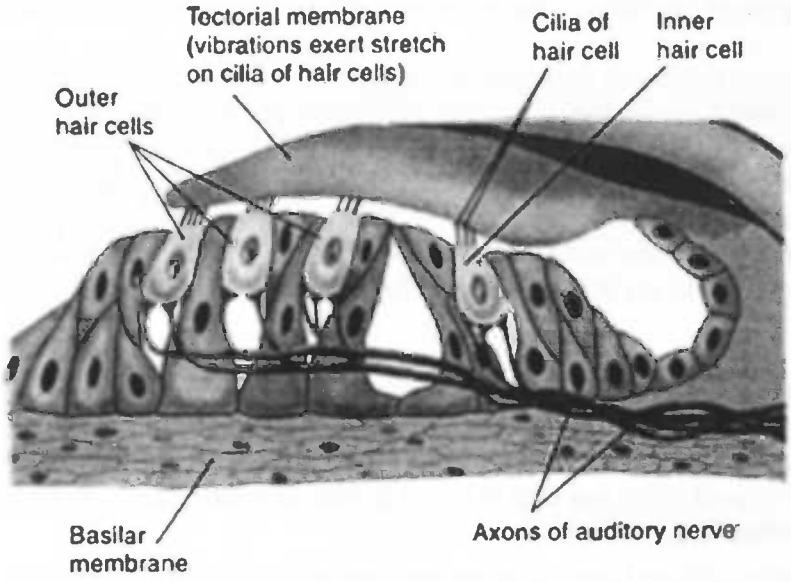


Figure 5. The basilar membrane with embedded hair cells.

1.4 The peripheral auditory system

Pressure waves produced by vocal chords, instruments or any other source, enter the ear, being the primary information for inducing a psycho acoustic perception of sound. The tympanic membrane conducts these pressure waves to the middle ear bones, which amplify the incoming signal and submit their energy through the oval window to the liquid filled cochlea. The latter contains the *organ of Corti*, with the basilar membrane (BM) (figure 5). In humans, about 3500 inner hair cells (IHC's) and 12000 outer hair cells (OHC's) are located upon the BM. Each IHC synapses with approximately 20 fibers in the cochlear nerve, a branch of the auditory nerve. In contrast, approximately 30 OHC's are innervated by a single fibre. It is currently believed that the inner hair cells are of primary importance in transmitting sound information to the brain, whereas the OHC's modulate the physical properties of the BM. The BM and IHC's are tonotopically organized: The IHC's in the outer cylinder of the curled-up cochlea respond to the highest frequencies, whereas closer to the centre of the cochlea, the IHC's respond to lower frequencies. Since ICH's only fire when they bend away from the oval window, and not towards, they transfer only half of the pressure wave input. This is termed half-wave rectification. IHC's are tuned to a small range of frequencies and respond maximally to their characteristic frequency (CF). The firing patterns of IHC's in response to sound stimuli are reflected in the firing pattern of adjacent fibres in the cochlear nerve, although the latter perform high pass filtering on their input (Narayan et al., 1998).

1.5 The central auditory system

Tonotopic organization is maintained throughout successive pathways in the AS and within the auditory cortex (AC). In the mammalian AS the cochlear nerve projects to several cochlear nuclei in the medulla (figure 6). From there, nerve fibers send axons to the controlateral nuclei of the *superior olivary complex,* which are believed to perform spatial processing. A large bundle of fibers, called the *lateral lemniscus* holds neurons from both the superior olivary complex and the cochlear nuclei and projects to the *inferior colliculus* (IC) in the midbrain. In the IC neurons are organized in columns, which are sensitive to a sharp frequency range. Within these columns rows of neurons are sensitive to specific periodicities. The IC sends axons to the *medial geniculate nucleus* in the thalamus. The latter consists of several nuclei, with yet to be identified functions, and projects to the auditory cortex.

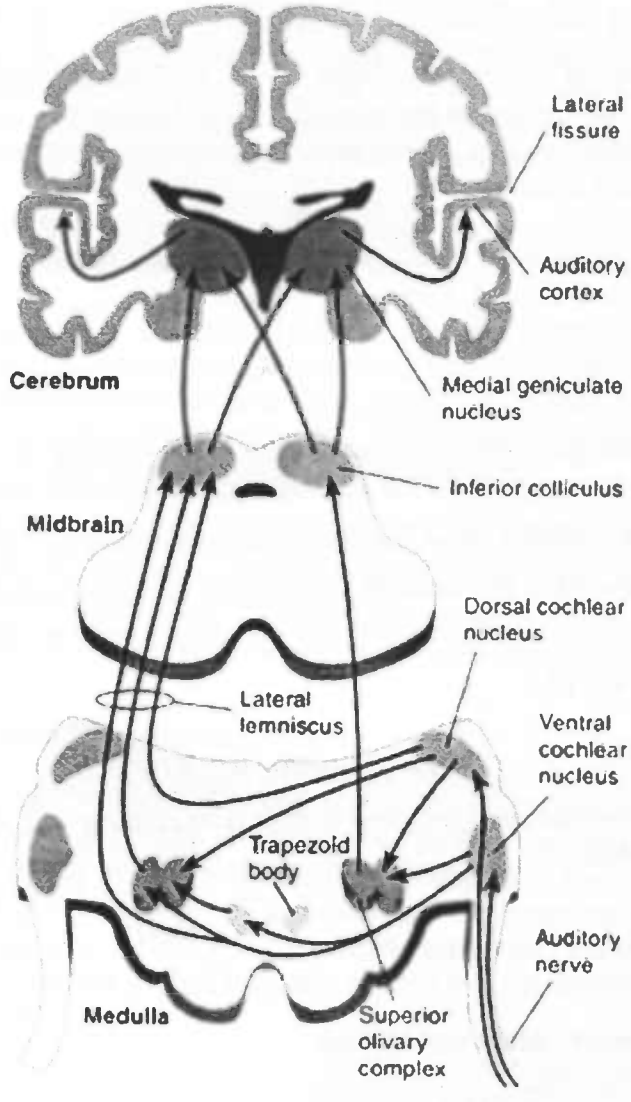


Figure 6. Pathways from auditory nerve to the audio cortex.

2. Methods

2.1 Associative networks

In a Delphi environment an associative network of 21 relaxation oscillators was simulated to test its capacities in supervised learning and one of 11 oscillators was simulated for testing the network under unsupervised conditions. Both networks were composed of fully interconnected Terman-Wang relaxation oscillators. The equations for these oscillators were given by:

$$\frac{dx}{dt} \quad f(x) \quad y \quad I \quad Y$$

$$\frac{dy}{dt} \quad H(g(x) \quad y)$$

where $f(x) = 3x - x^3 + 2$ and $g(x) = \Delta\,(1 + \tanh(x/\mathrm{E}))$

I represents the sum of lateral and external input to the oscillator, whereas Δ and E are parameters that determine the oscillation envelope and His necessary to achieve oscillatory state. Some pseudorandom noise was added by Y for facilitating desynchronisation.

For supervised training a supervised hebbian learning rule was used, according to:

$$w_{ij} \quad \frac{1}{aN} \sum_{v} (\int_{i}^{v} \quad a)(\int_{j}^{v} \quad a)$$

where w_{ij} is the synaptic weight from oscillator k to i, N is the number of patterns, v is the pattern index and a is the average probability of a vector bit being equal to "1". Here a constant value is used as an approximation. $\int_{i}^{v}$ is the set of patterns to be stored in the network. Total input for each oscillator is given by:

$$I_i \quad E_i \quad \sum w_{ij} \varsigma\,(x_j)$$

Here, E_i is the external input.

Weight updating during unsupervised learning was achieved according to the following differential equation:

$$\frac{dw_{ij}}{dt} \quad \Im\,(x[t,i])(\quad w \quad M\,(x[t,j]))$$

For the unsupervised network, total input follows:

$$I_i \quad E_i \quad \sum \frac{w_{ij} \quad w_{inh}}{a_2 N} \varsigma\,(x_j)$$

Where N is the number of oscillators in the network. A static inhibitory weight w_{inh} is used, similar to (Lourenco, 2000). This means that if the excitatory connection strength between two oscillators is insufficient, inhibition wil take over and cause desynchronization.
This learning rule incorporates a simple version of Hebbian synaptic plasticity, in this case gradually enforcing connections from active neurons when the neuron itself is active, and diminishing connections from inactive neurons. The weighted input is scaled in adjustment to the network size.
A more detailed discussion of the unsupervised learning rule can be found in (Marshall, 1995).

$\sigma(x)$ is a sigmoid function: $1/(1 + \exp(-K*(x - T)))$

K is a parameter to adjust the steepness of the function, and T is a threshold parameter.
The basic setup of the experiment was derived from Lourenco et al. (2000), with the major difference being the kind of oscillators used. Lourenco used bursting oscillators instead of relaxation oscillators. In addition, to achieve different network behaviours, Lourenco adopted various parameters, whereas parameters in the current model remained constant for many situations.

The network for timbre recognition was constructed with 200 fully interconnected neurons and was trained supervised with segregated sound patterns of two different instruments. After training, test patterns from the two instruments at different pitches were presented to test the model's capacity identifying the correct instrument. During this testing the synaptic weights between neurons were fixed, so additional learning could not take place.
For the numerical integration, Euler was used with a stepsize of 0.1. Validity of results was verified at smaller stepsizes.

Parameters used, except where mentioned:
common parameters: H= 0.02, Δ = 6.0, E = 0.1, K = 50
supervised model: T = -0.5, a = 0.38
unsupervised model: a_2 = 1.3, M= 0.8, T = 1.1, ϑ= 0.005, w_{inh} = 0.65

2.2 Audio patterns

Samples at a variety of pitches from cello, guitar, harp, flute and saxophone served as input for the model. These samples came from a professional sample database on the internet (http://soleil.ircam.fr) and were mono sampled at 16khz. Patterns generated from these samples by the segregation network, were binary and their similarity was expressed in hamming distance (see also: table 1). In the timbre recognition experiment the tuba was tested against the guitar, the harp against the saxophone and the cello against the flute.

2.3 Inner hair cells

To simulate the IHC's the routines *MakeERBFilters*, *ERBFilterBank*, and *MeddisHairCell* from the Auditory Toolbox 2.0 for Matlab were used. These routines together imitate processing by the cochlea (Meddis and Hewitt, 1991). *MakeERBFilters* was used to set the frequency range for the cochlea, the number of frequency channels and their individual frequency range. For this simulation 64 channels were used in a cochlear frequency range of 50 to 8000 Hz. Note that, in line with the nonlinear processing by the cochlea, the frequency range of each channel differed and that there was some overlap between channels. *ERBFilterBank* was used to distribute the audio input among the frequency channels whereas the *MeddisHairCel* routine performed half-wave rectification and low pass filtering upon the data per frequency channel and generated a spike output when an action potential threshold was exceeded. The output values of the simulation were binary and stored in a matrix in a text file (frequency x time), for use in Delphi. Each spike corresponded to the length of a sample unit in the wave file.

2.4 Segregation, autocorrelation and normalization

Timbre segregation was performed in Delphi. Data from each IHC was presented to an array of 150 delay loops, with a length range of 1 to 150 spikes. Expressed in time, the loops could deal with periodic patterns ranging from approximately 0,06 to 9,38 ms[1]. A build-up rule for coincidence detection was implemented according to the following algorithm:

$$\min(X(CF,t), H(\tau,t)) + 0.1*\mathrm{abs}(X(CF,t) - H(\tau,t))$$

The minimum value of a coinciding input spike (X) and a re-entering spike (H) at moment *t* plus their absolute difference times a factor 0,1 was used as a new input value for the delay loop. The loop, which amplified the F_0, was determined and for all the frequency channels loops of the same length the output was used for autocorrelation. Their two nearest neighbours were considered as well. The other 147 loops were left out, since they were assumed to provide only little extra periodic data, whereas they enlarged rendering time enormously.

Due to the build-up rule the output of the delay loops was analogue. This output was subjected to autocorrelation by Licklider autocorrelators. The 192 (3 loops x 64 channels) output patterns were autocorrelated independently. Autocorrelation on the segregated data was performed over a time frame of 1000 samples. The shortest and longest interval to be detected were set to 1 and 200 samples respectively. Autocorrelation was calculated according to the following equation:

$$A(t,\Omega) \quad \int_{k\ 0}^{K\ 1} r(t\quad k)r(t\quad k\quad \Omega)$$

[1] The duration of a spike was 1/16000 (Hz)) = 0,06 ms. Thus, a delay loop with length 150 could deal with periodicities of 150*0,06 ms.

Where t is the time step in a time frame of K time steps, τ is the interval length and $r(t)$ is the spike probability at time t. The resulting interval patterns for a single time frame were integrated per interval, that is, all the values in a matrix of 1000 time steps and 192 autocorrelators were all added up, so that the final outcome was a set of 200 analogue interval values. These spike interval distributions were scaled to a standard interval length, using a simple linear filtering sampling scheme with a scaling factor calculated from the quotient of the length of the delayloop and a standard length of 200.
This resulted in a periodicity distribution normalized for the pitch, as estimated by the delayloop circuit, for further processing.
By means of lateral inhibition the contrast between these values was enhanced to obtain sharper peak patterns. The application of lateral inhibition served to bring important intervals to the foreground. The following formula was used:

$$x_i \quad \overset{j\ i\ d}{\underset{j\ i\ d}{f}} q_j(1.4 \quad \frac{|j \quad i|}{2.72})$$

Where qj is the output of the normalised correlogram obtained in the previous step, and d is the width of the lateral inhibitory filter, set to 7 here. By means of lateral inhibition the contrast between these values was enhanced, that is, high values inhibited their lower neighbour values. Finally, the analogue values were compared to a constant threshold to obtain binary patterns.

3. Results

3.1 Terman-Wang oscillator network (supervised)

The first test case for the supervised network consisted in retrieving previously acquired binary patterns. Three non-overlapping patterns were taught during the training stage. During the retrieval stage hebbian learning was disabled and the acquired patterns were presented to the network simultaneously. As can be seen in figure 7, the three memorized patterns were desynchronised after one period.

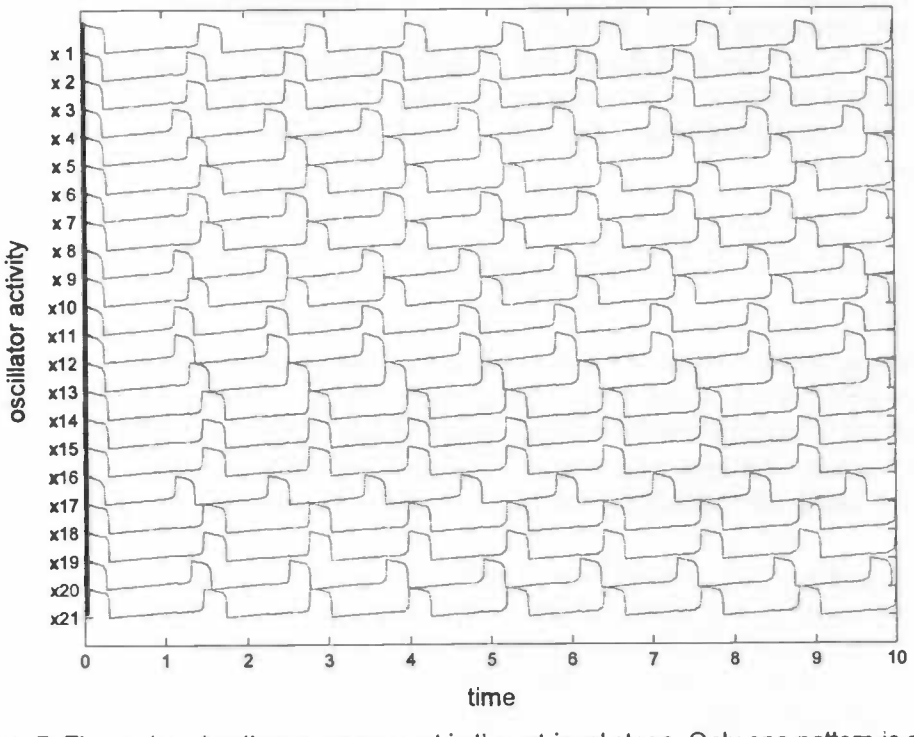


Figure 7. Three stored patterns are present in the retrieval stage. Only one pattern is active at a time. Patterns are: (0,1,1,0,1,0,1,0,0,1,0,0,1,0,0,0,0,0,0,1,0), (0,0,0,1,0,0,0,0,1,0,1,1,0,0, 0,0,1,0,0,0,0) and (1,0,0,0,0,1,0,1,0,0,0,0,0,1,1,1,0,1,1,0,1). Thick blue lines indicate input.

In a second experiment the network's capacity for pattern completion was tested. The network was supervised trained on one large and three small binary patterns, which had no overlap: (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0), (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0), (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0) and (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1). Then, during the retrieval stage, three degraded version of the large memorized pattern were presented: (0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0), (0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,0,0,0) and (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0). The results of these simulations are depicted respectively in the figures 8,9 and 10. Note that the first test pattern was only little degraded and that the original pattern was completed. The second test pattern was further degraded and as a consequence completion of the original pattern was ambivalent. The third pattern had so much degradation, so that no pattern completion occurred.

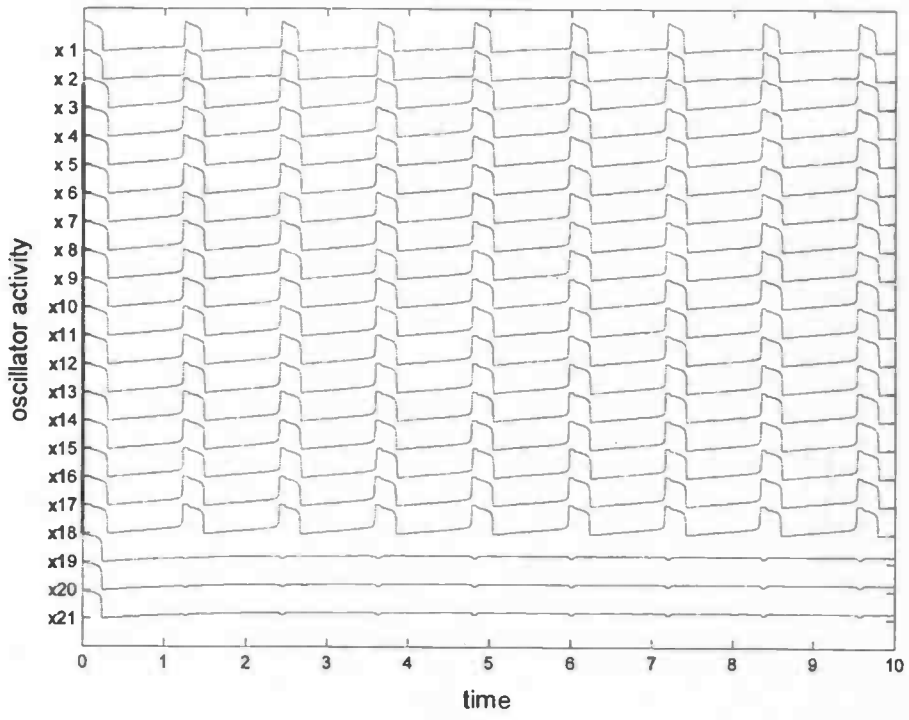


Figure 8. Pattern (0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0) is presented to the network and the original pattern is retrieved. Here, as in figures 8 and 9, T

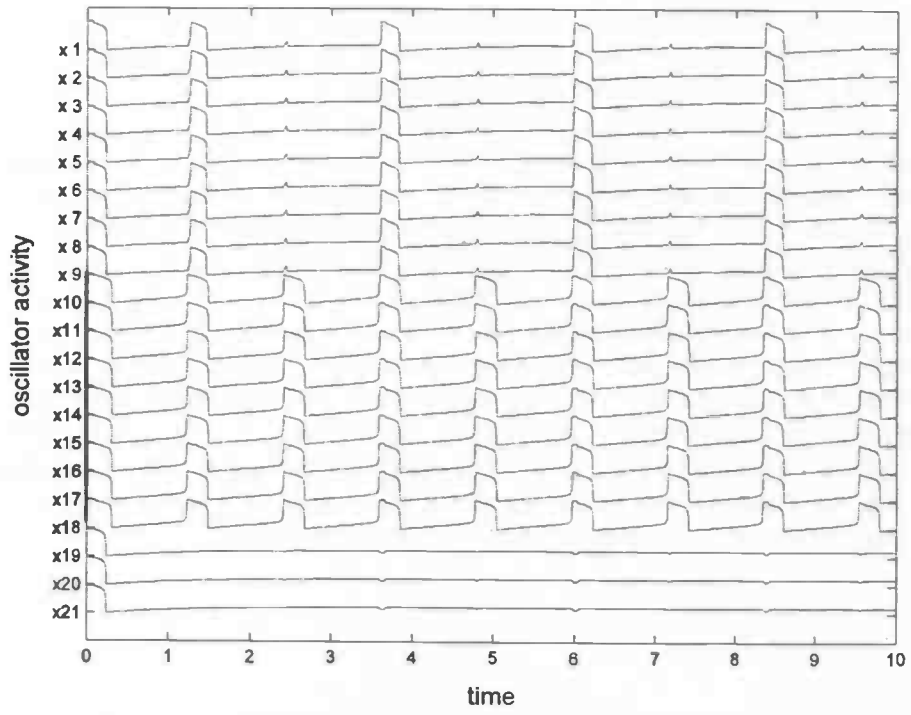


Figure 9. Pattern (0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,0,0,0) is presented to the network and the original pattern is completed in half of the oscillation periods.

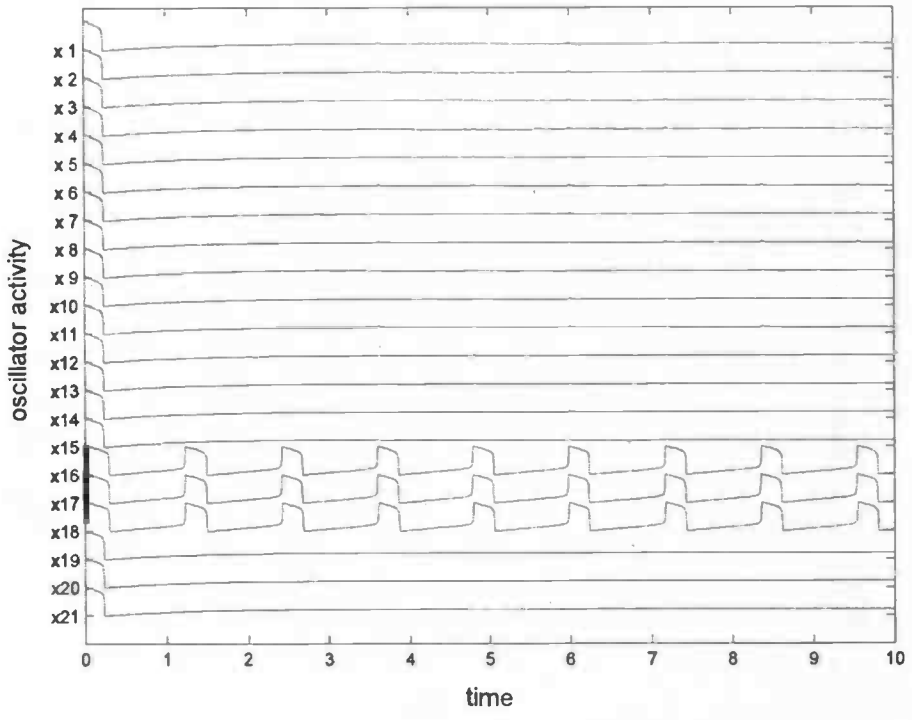


Figure 10. Pattern (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0) is presented to the network. Now, only the oscillators that receive input become active.

During a third experiment the same network with the four memorized patterns was tested with analogue input patterns. First, an analogue version of the large test pattern was presented to the network (0.1,0.1,0.1,0.1,0.1, 0.1,0.15,0.15, 0.15,0.15,0.15,0.15,0.2,0.2,0.2,0.2,0.2,0.2). Note that, corresponding to the input values, this pattern can be considered as three sub-patterns. As a result, the network segmented three smaller patterns instead of completing the original pattern (figure 11). For another analogue completion test, the network was first supervised trained to segment three 6-bit patterns and then presented with three equal sized analogue patterns of the form (0.065,0.065,0.14,,0.14, 0.27,0.27). The binary patterns' configuration did not correspond to that of the analogue patterns. As can be seen in figure 12, at an early stage, patterns were segmented according to their input value. However after a few periods the previously memorized patterns were segmented. It is interesting to note that in both cases of analogue input the distance between spikes correlates with the input values: the larger the input the smaller the spike distance. This can be interpreted as analogue coding in spike distance.

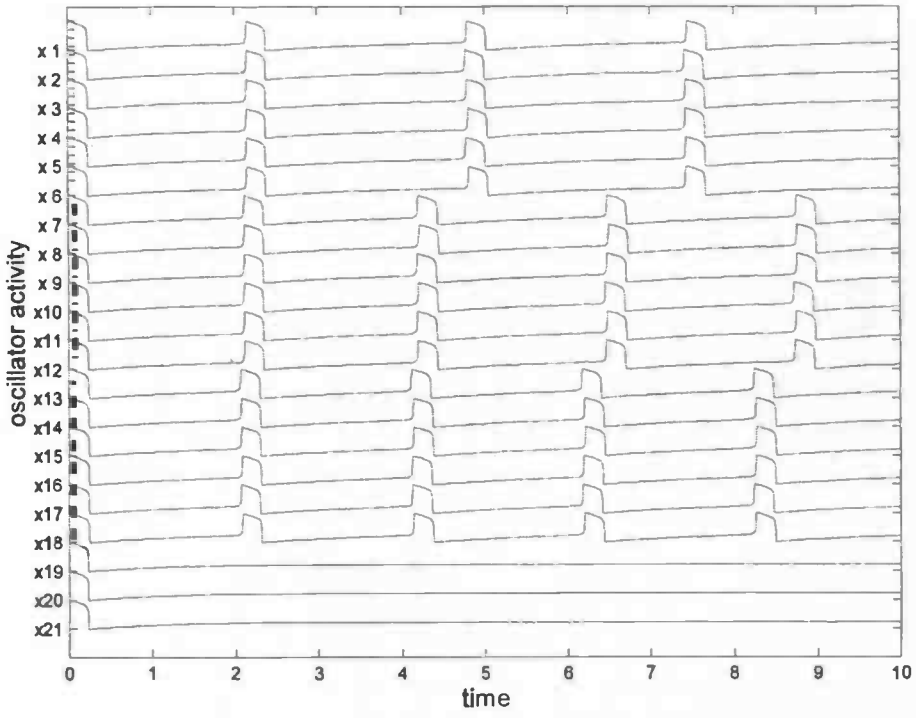


Figure 11. Completion of an analogue pattern. Three levels of input (0.1, 0.15, 0.2) results in clustering of six sub patterns

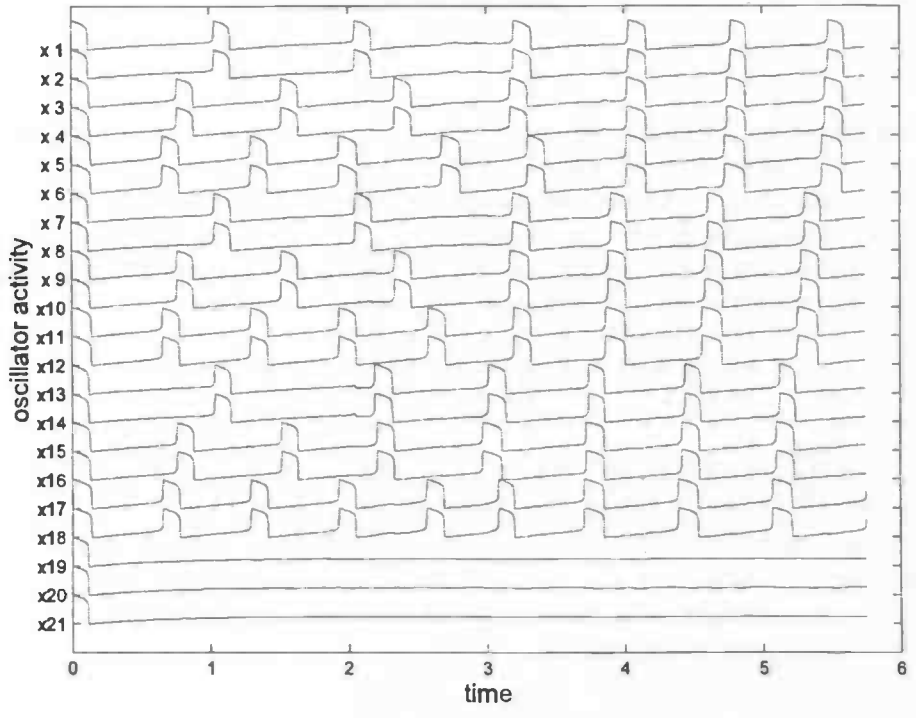


Figure 12. Processing of analogue patterns. The network memorized three 6-bit patterns and is stimulated with three analogue patterns (different combination from the binary pattern). First, the three analogue patterns are segmented, but after a few oscillations the binary patterns take over.

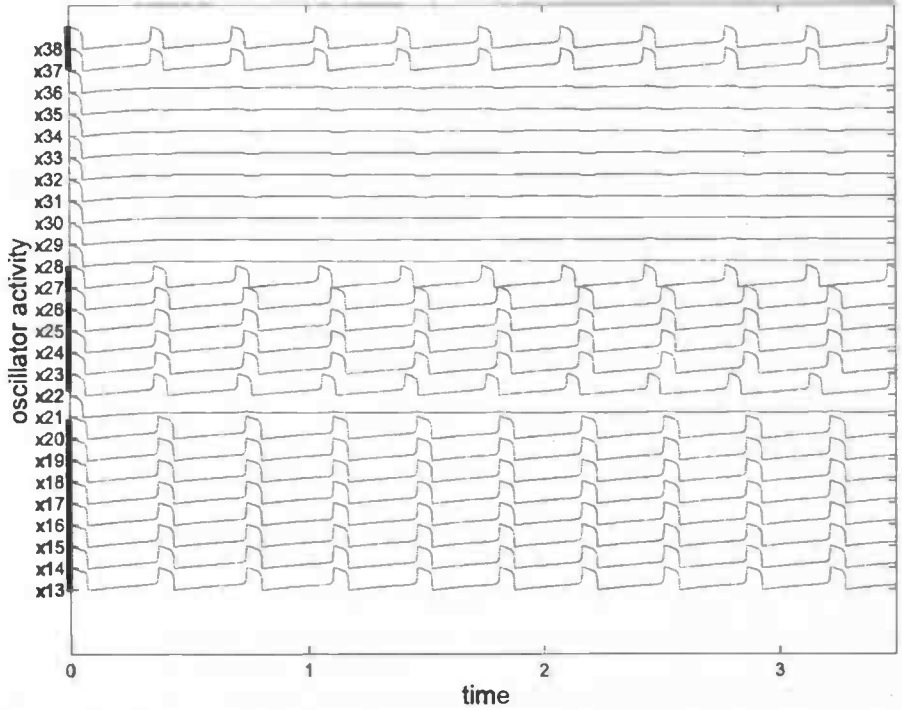


Figure 13. The network memorizes 7 strongly overlapping patterns and is presented with one of them, while in addition four other oscillators are triggered as well. Nonetheless, the memorized pattern is successfully completed in a few periods. b=0.6

For a last supervised test an important situation not present in the simulations presented by Lourenco et al. was considered, that is, the retrieval of patterns when large overlap exists between stored segments. The same set of parameters was applied, with only b set to 0,6. This served to decrease the number of active segments, clarifying the picture. A network of 38 oscillators was designed and trained with the following seven patterns:
(1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1),
(1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,0,0,0,0,0,0,0),
(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,0,0,0,0),
(0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0,0,0),
(0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
(1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1),
and (0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1).
During a segmentation test it was presented with the fifth pattern and in addition, input to oscillators x22, x27, x37 and x38. In spite of the extra input, pattern five was completed successfully, as the other four oscillators were desynchronised after a few periods.

3.2 Terman-Wang oscillator network (unsupervised)

To test unsupervised learning a relaxation network of 11 oscillators was built. At the beginning patterns were trained supervised, but at the experimental stage the network was equipped with an unsupervised Hebbian learning rule. In a first experiment the network's ability to memorize new patterns was explored. The network was supervised trained with a 5-bit and a 6-bit pattern and in the unsupervised configuration it was presented with the 6-bit pattern and a 2-bit pattern. Figure 14 depicts the results: Presenting a 2-bit pattern resulted in segmentation of the original 5-bit pattern into two smaller patterns. To test the long-term memory effect of this segmentation a 5-bit pattern was presented to the network (figure 15). At first the 5-bit pattern was active as a whole, but after a few oscillations it desynchronised back into a 2-bit and a 3-bit pattern. Thus, memorization of the two sub-patterns had occurred successfully.

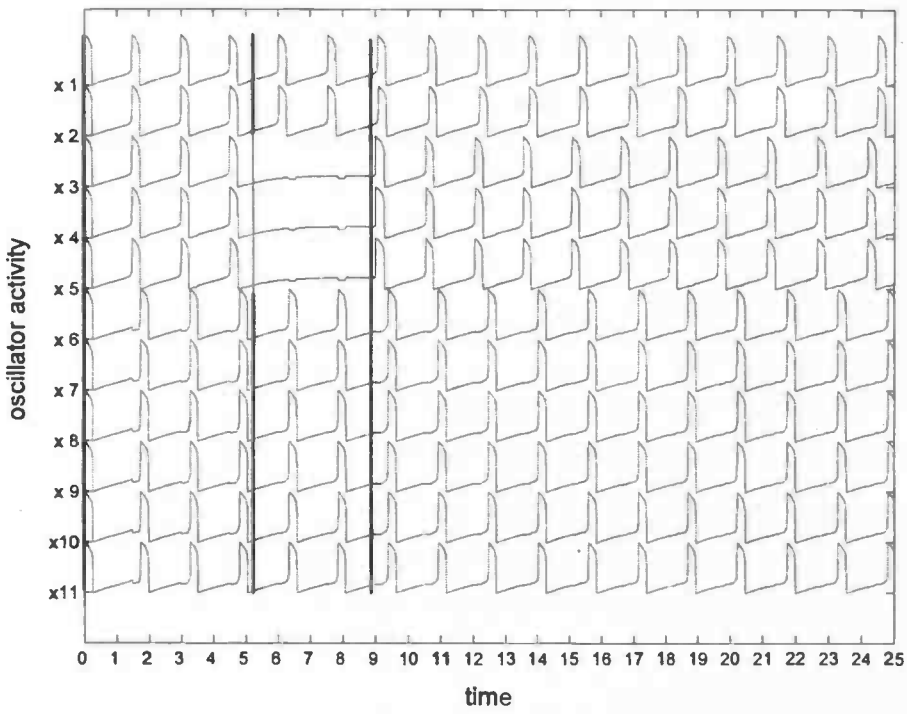


Figure 14. The network, which memorizes a 5-bit and a 6-bit pattern is shortly presented with a 6-bit and a 2-bit pattern. As a consequence the 5-bit pattern is segmented into two smaller patterns. Thick blue lines indicate input; thin blue lines indicate no input.

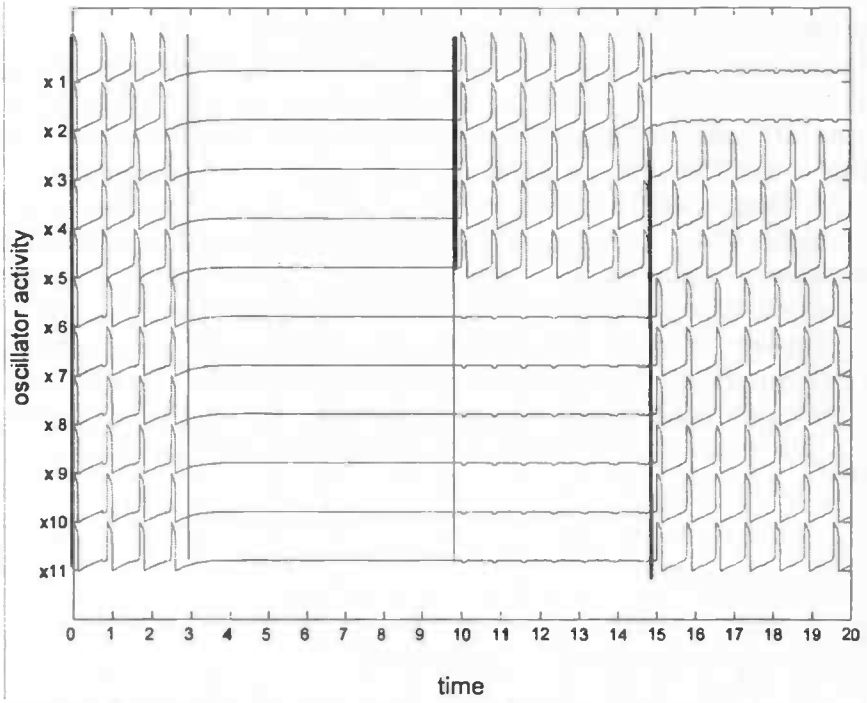


Figure 15. Long-term memory effects are visible after stimulation with a 5-bit pattern, as it segments into a 2-bit and a 3-pattern.

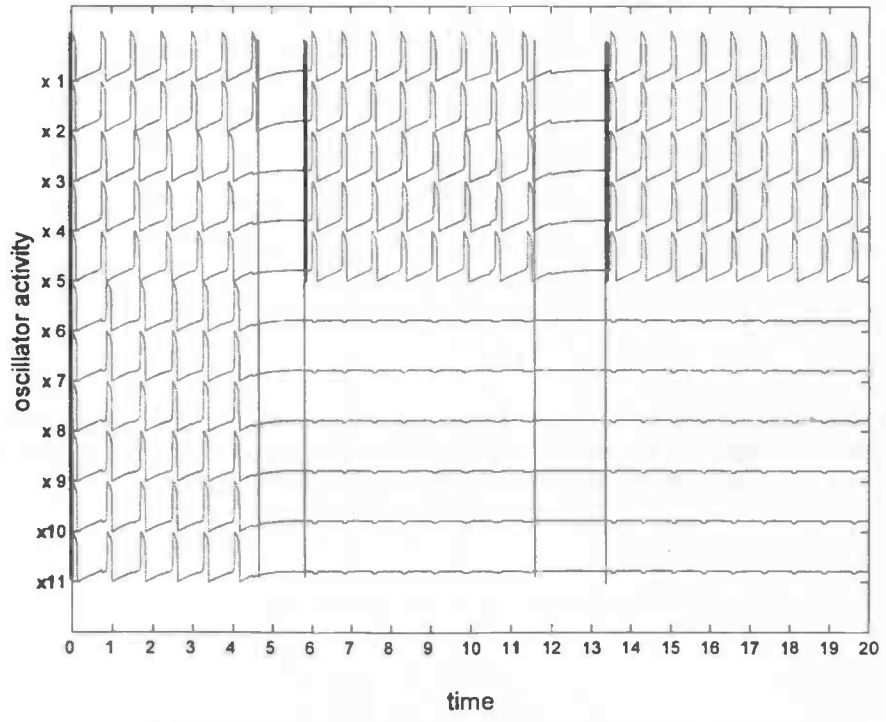


Figure 16. Successful rebinding of a previously segmented 5-bit pattern.

It was also tested whether the network was able to 'rebind' segmented patterns. In order to do so the experiment in figure 15 was repeated. However, this time the 5-bit pattern was presented a third time. After the second presentation, the 2-bit and the 3-bit pattern were still present, but after a third presentation they re-binded into a single 5-bit pattern.

To test whether segmentation would occur in any situation wherein a new pattern was presented, the above network (which memorized the 2-bit, 3-bit and 6-bit pattern) was presented with a 5-bit pattern at oscillators $x6$, $x7$, $x8$, $x9$ and $x10$ and a pattern at oscillator $x11$. The last oscillator was considered as independent pattern, since it had an input value, which was 0.1% degraded as compared to the other input values. The two patterns were presented to the network for a short moment and. Although oscillator $x11$ was desynchronised for a moment, a 6-bit pattern was completed after a few oscillations (figure 17). However, in a second trial the two patterns were presented longer and this resulted in sustained desynchronisation of oscillator $x11$ (figure 18).

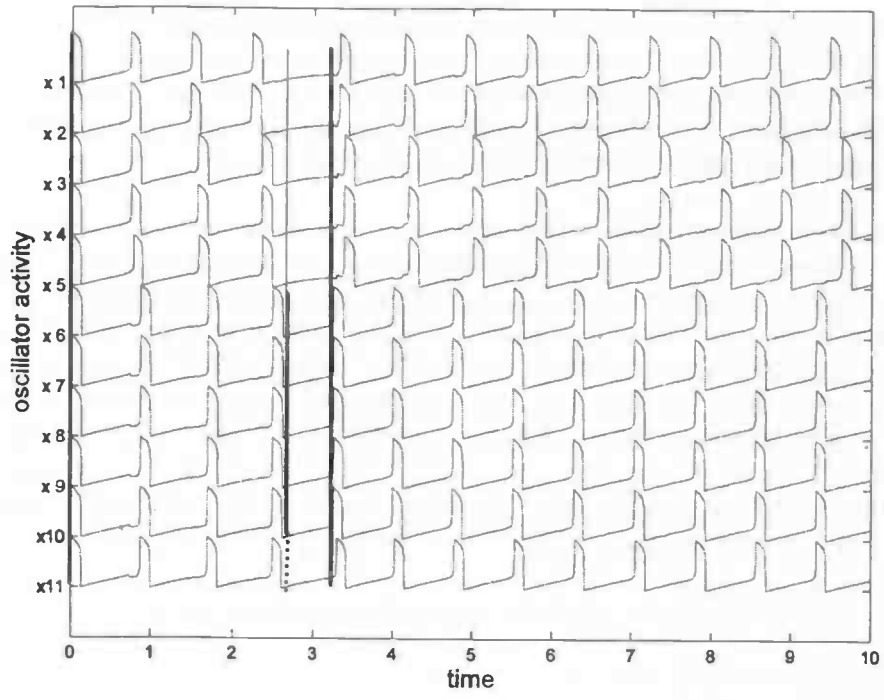


Figure 17. For a short moment a 5-bit pattern is presented, joined with a 1-bit pattern, which has a smaller input value. The latter desynchronises shortly, but soon resynchronises with the 5-bit pattern.

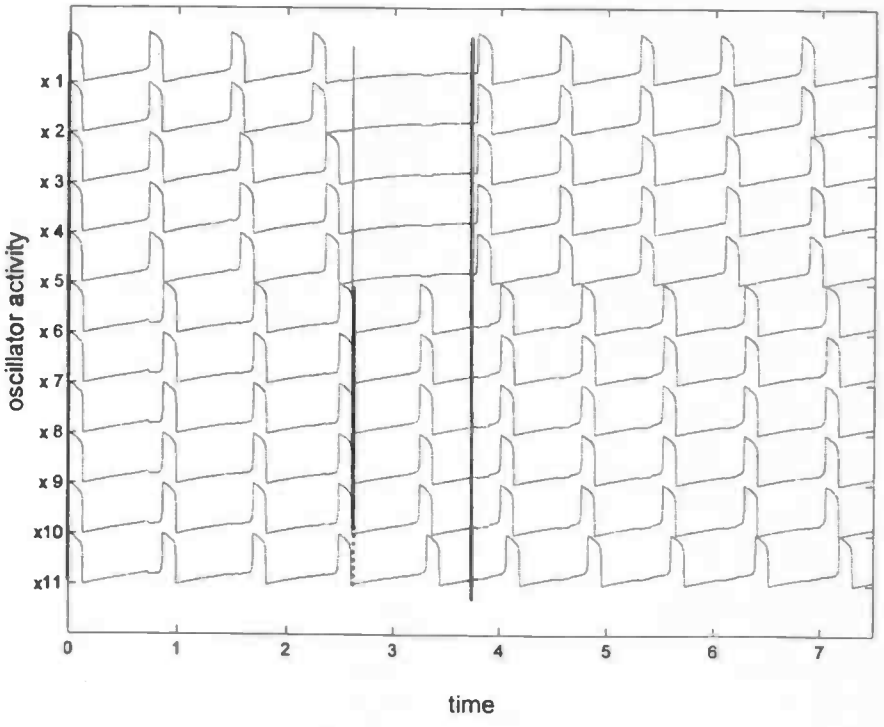


Figure 18. The experiment in figure 13 is repeated, but now the weakened stimulation occurs longer. As a result oscillator *x11* remains out of synchrony with the rest.

3.3 Segregated patterns

Between all the segregated audio patterns the hamming distance (HD) was calculated, averaged per instrument group and then compared between groups (see table 1).

Except for the cello and flute, the instruments displayed the lowest average hamming distance (AHD) with themselves. Guitar and harp had the lowest within-group AHD and in addition the lowest between-groups AHD with one another. The tuba had a low within group AHD and differed clearly from other groups, in particular from the string instruments. The sax was closest related to the flute. Calculation of AHD's for the classes string instruments and wind instruments revealed a small AHD for strings (46,89) as compared to the overall average (74,25).

Table 2 depicts AHD's between original samples and their transposed variants. In general these AHD's were much lower than the AHD's for whole instrument groups. The cello at tone E5 and the guitar at B4 were exceptions.

	Cello	Guitar	Harp	Tuba	Sax	Flute	
Cello	61	58	64	101	83	71	
Guitar	58	22	33	107	78	70	
Harp	64	33	29	120	81	74	
Tuba	101	107	120	40	88	90	
Sax	83	78	81	88	68	73	
Flute	71	70	74	90	73	71	
Average			46,89			75,67	74,25

Table 1. AHD between instrument groups. The bottom row respectively depicts average values for strings, winds and the overall average.

Cello	61	Guitar	22	Harp	29	Flute	71	Sax	68	Tuba	40
D, B4	57	G, B4	30	B4	4	B4	43	B4	22	B2	14
D, C4	27	G, C4	5	C4	7	C4	18	E4	45	D3	13
D, D4	1	G, D4	12	E4	5	D4	12	D4	31		
D, E4	35	G, E4	9	F4	9	E4	30				
D, C5	21	G, C5	8	G4	7	G5	45				
A, C4	17	D, C4	10								
A, E4	36	D, D4	7								
A, E5	65	B, C4	8								

Table 2. AHD between original and transposed versions of the same sample. Left columns display the tones, right columns the AHD.

The same procedure was repeated for a subgroup of samples. However, this time, 5% or 20% noise was added, or loudness was reduced with 10% or 50%. The AHD's are depicted in table 3. In this case, for each instrument the within-group AHD was larger then the between-groups AHD. Moreover, the instrument class AHD's (34,78 and 65,5) were much smaller then the overall AHD of 88,11.

	Cello	Guitar	Harp	Tuba	Sax	Flute	
Cello	39	44	53	94	93	69	
Guitar	44	18	25	102	102	75	
Harp	53	25	12	104	80	74	
Tuba	94	102	104	43	91	81	
Sax	93	102	80	91	36	61	
Flute	69	75	74	81	61	43	
			34,78			65,33	88,11

Table 3. AHD between instrument groups, of samples with noise added, or reduced loudness. The bottom row respectively depicts average values for strings, winds and the overall average.

Finally, AHD's for the different distortions were calculated to analyse their effects. Table 4 represents typical AHD's for these situations. Reduced loudness of 10 and 50% provided the same results. These were not different form the original samples (data not shown). There was in few cases a small effect of 5% noise and there was a somewhat clearer effect of 20% noise.

	AHD
10% reduced loudness	69,4
50% reduced loudness	69,4
20% noise	70,1
5% noise	69,4

Table 4. AHD for the guitar B-string at C4 as compared to the other samples at different levels of reduced loudness or noise.

3.4 Audio-pattern classification

The oscillator network was tested three times with pairs of samples, which were: tuba-guitar, saxophone-harp and cello-flute. According to *ad hoc* definition, a sample was classified correctly when it was in more than 50 of the time steps classified as the expected instrument, and never as the unexpected one. The flute-cello test resulted in six correct flute classifications out of 15. Of the cello none of the 21 samples was correctly classified (table 5). The binary pattern of the trained flute contained larger active pattern size (the number of '1' bits) than the cello.

	Flute C4	Cello C4
Flute, C4	27	0
Flute, C4 1up	283	0
Flute, C4 2up	278	0
Flute, D4	0	61
Flute, D4 1up	11	54
Flute, D4 2up	4	43
Flute, E4	0	0
Flute, E4 1up	0	53
Flute, E4 2up	78	0
Flute, B4	168	0
Flute, B4 1up	14	0
Flute, B4 2up	6	2
Flute, G5	98	0
Flute, G5 1up	136	0
Flute, G5 2p	217	0

Average HD between flutes and cellos: 71
HD between trained flute and cello: 53
Flute: 129 x 1, 71 x 0
Cello: 90 x 1, 110 x 0

	Flute C4	Cello C4
Cello (D), C4	290	113
Cello (D), C4 1up	32	0
Cello (D), C4 2up	71	0
Cello (D) D4	6	0
Cello (D), D4 1up	7	0
Cello (D), D4 2up	6	0
Cello (D), E4	58	0
Cello (D), E4 1up	1	13
Cello (D), E4 2up	0	43
Cello (D), B4	6	6
Cello (D), B4 1up	6	0
Cello (D), B4 2up	230	0
Cello (D), C5	9	189
Cello (D), C5 1up	26	0
Cello (A), C4	0	0
Cello (A), C4 1up	21	0
Cello (A), C4 2up	0	0
Cello (A), E4	6	0
Cello (A), E4 1up	1	0
Cello (A), E4 2up	6	0
Cello (A), E5	0	0

Table 5. Classification results for flute-cello. Names in the upper row indicate the trained samples. Left columns indicate the tested samples, where 1up and 2up denote transpositions and characters between brackets denote the string type of a string instrument. Middle and right columns represent the number of classifications. Bold names indicate correct classifications.

	Guitar C4	Tuba D3
Tuba, B2 1up	0	932
Tuba, B2 2up	0	939
Tuba, D3	0	1340
Tuba, D3 1up	0	1081
Tuba, D3 2p	0	1037

Average HD between tubas and guitars: 107
HD between trained tuba and guitar: 137

Guitar: 59 x 1, 141 x 0
Tuba: 139 x 1, 61 x 0

	Guitar C4	Tuba D3
Guitar (G), C4	67	15
Guitar (G), C4 1up	0	12
Guitar (G), C4 2up	0	7
Guitar (G), C5	0	7
Guitar (G), C5 1up	0	7
Guitar (G), C5 2up	0	13
Guitar (G), D4	0	7
Guitar (G), D4 1up	0	6
Guitar (G), D4 2up	0	6
Guitar (G), E4	0	6
Guitar (G), E4 1up	0	6
Guitar (G), E4 2up	0	6
Guitar (G), B4	0	162
Guitar (G), B4 1up	0	41
Guitar (D), C4	0	7
Guitar (D), C4 1up	0	5
Guitar (D), C4 2up	0	6
Guitar (D), D4	0	6
Guitar (D), D4 1up	0	6
Guitar (D), D4 2up	0	6
Guitar (B), C4	0	63
Guitar (B), C4 1up	0	44
Guitar (B), C4 2up	0	6

Table 6. Classification results for tuba-guitar.

Testing the tuba against the guitar resulted in correct classification of all tubas (table 6). In contrast, guitars were never classified as a guitar, but twice as a tuba. The binary pattern of the trained tuba contained a larger active bit size than the guitar did and their HD was large: 137.

The saxophone was tested against the harp (table 7). In this situation samples with added noise or reduced loudness were tested as well. Except for one, all of the saxophones were classified correctly but several harps were also classified as saxophones. For the harp noise and reduced loudness had negligible effect, but for saxophones this could lead to large differences between four adaptations. Again, the larger binary pattern (the saxophone) biased the classification.

	Sax D4	Harp C4
Sax, E4	552	0
Sax, E4, L10	563	0
Sax, E4, L50	558	0
Sax, E4, N5	491	0
Sax, E4, N20	320	0
Sax, E4 1up	283	0
Sax, E4 1up L10	284	0
Sax, E4 1up L50	282	0
Sax, E4 1up N5	165	0
Sax, E4 1up N20	325	0
Sax, E4, 2up	570	0
Sax, D4	1154	0
Sax, D4 L10	1154	0
Sax, D4 L50	1153	0
Sax, D4 N5	757	0
Sax, D4 N20	416	0
Sax, D4 1up	431	0
Sax, D4 2up	747	0
Sax, B4	173	0
Sax, B4 1up	5	0
Sax, B4 2up	219	0

Average HD between harp and sax: 81
HD between trained harp and sax: 64
Sax: 126 x 1, 074 x 0
Harp: 78 x 1, 122 x 0

	Sax D4	Harp C4
Harp, E4	22	0
Harp, E4 L10	20	0
Harp, E4 L50	23	0
Harp, E4 N5	24	0
Harp, E4 N20	21	0
Harp, 1up E4	60	0
Harp, E4 1up L10	60	0
Harp, E4 1up L50	59	0
Harp, E4 1up N5	66	0
Harp, E4 1up N20	67	0
Harp, 2up	67	0
Harp, F4	30	0
Harp, F4 L10	4	0
Harp, F4 L50	3	0
Harp, F4 N5	3	0
Harp, F4 N20	3	0
Harp, F4 1up	6	0
Harp, F4 2up	23	0
Harp, G4	4	0
Harp, G4 1up	3	0
Harp, G4 2up	3	0
Harp, B4	16	0
Harp, B4 1up	15	0
Harp, B4 2up	5	0
Harp, C4	180	117
Harp, C4 1up	3	0
Harp, C4 2up	4	0

Table 7. Classification results for sax-harp. LN10 and LN50 respectively denote volume reduction of 10% and 50%. N5 and N20 respectively denote 5% and 20% noise.

4. Discussion

4.1 Average hamming distances

For the samples without reduced loudness or added noise there was a great variation in AHD between samples of the same instrument, so that the average distance within an instrument group could even be larger then those between groups. This was the case for cello and flute. Analysis of the effect of transposition revealed that the distance between an original sample and its transposed versions was generally small. Therefore the large average group AHD must be ascribed to the differences between sampled sounds. Although AHD's were often low between samples and their transpositions, a few were unexpectedly large. Calculations for only the strongest amplified delay loops and its neighbours probably neglected too much data and might account for these large differences. It should be noted in addition that there were no significant correlations between tone distances and hamming distances between samples.

The large hamming differences found within instrument groups suggest that periodicity alone cannot account for timbre recognition. The AHD's from the set with samples that had noise added or a reduced loudness provided results that were more in line with the expectations that periodicity provides sufficient information for the identification of instruments. However, this sample set contained two samples of each instrument at different transpositions and adaptations. Since neither of these showed large effects on the AHD, the small within-group AHD's were mainly caused by the small number of original samples used. Nonetheless, AHD's within the string and wind classes appeared smaller than AHD's between them, thus it is possible that periodicity actually does provide some cues for sound recognition.

4.2 Sound recognition

The oscillation network displayed biased classification in direction of the sound patterns with the largest active ('1') bit pattern size throughout the three tests. This was probably mainly due to the setting of the α parameter. A higher value for α results in a bias toward larger patterns. Both tuba/guitar and sax/harp displayed strong completion of one instrument. There were effects of the active input size present in the patterns but there were probably also effects of HD's between the trained instruments. There is however too little data to confirm such an assumption. In order to investigate specific variables it is necessary to use task-specific manipulated patterns.

4.3 Stability of oscillator dynamics

The interaction of many time dependent factors in the associative oscillator network makes it capable of very complex behaviour, which can be hard to interpret and/or adjust. The results in 3.1 clearly show the network is capable of performing well under a host of conditions. However, certain general restrictions apply to oscillator networks, which also affect this model.

One such restriction is known as segmentation capacity. Cairns, Baddeley and Smith (1996) found that there is a limit to the number of stable phases one can expect a system of interacting oscillators to maintain, and that this limit is fairly low. In other words, coupled oscillators will show a tendency toward global synchronization when many segments are activated. This is one of the reasons heavy overlap between segments significantly weakens the ability of the network to perform successful recognition.

In addition, the simple Hebbian learning rule used here is far from optimal in most situations.

Several methods for reducing pattern overlap can be devised. Complement coding is one way of doing this. It involves presenting the inverse of every input value along with the original, in effect normalising the pattern. However, this would require significant added complexity in the pre-processing stage, from a biological point of view.

The problem of reduced stability of recognition clearly presents itself in the results with the 200-oscillator network processing audio data. In agreement with the idea that pattern overlap may cause inaccurate recognition, a correlation between hamming distance of stored patterns and retrieval performance appears to be evident in table 5, 6 and 7.

The patterns generated in the pre-processing stage are representations of periodicity data in the audio signal. Specifically, peaks in the periodicity interval distribution are contrasted and thresholded, giving a binary pattern where the width of these peaks is translated to contiguous series of '1' bits. This results in patterns with a few long homogeneous bit series, causing a degree of overlap between patterns, independent of the dissimilarity of the periodicity distributions that generated these patterns.

Ideally, a peripheral processing stage rendering timbre representations would encode only position and relative amplitude of the periodicity peaks, minimizing overlap between dissimilar patterns.

4.4 Pattern segmentation versus rebinding

A key property of an unsupervised learning system is its ability to make decisions whether to store presented information or retrieve similar information from memory. In general, one would expect such a system to classify an input as an existing memory pattern when it is sufficiently similar. Conversely, if an unfamiliar pattern is presented repeatedly, it should be stored in memory.

The Hebbian laws of synaptic change applied in this research make storing of patterns dependent on covariances of oscillator activity. Thus, network dynamics and state determine whether links strengthen, weaken or remain unchanged. The state of the network presented in this paper is influenced both by its past dynamics, as well as the external input.

In contrast with the models proposed by Lourenco et al. (2000) and Wang (1990), the associative Terman-Wang oscillator network presented here shows a certain amount of stimulus-locking on very short timescales. That is, desynchronisation between simultaneously activated segments is rapidly achieved, but not instantly.

In a randomly connected network, this causes a short transient of synchronous activity in activated neurons. As Hebbian learning can act on this correlated activity, excitatory binding between oscillators can thus be induced by repeated simultaneous stimulation. Breaking of excitatory links is also achieved in a Hebbian fashion: during periods when an oscillator is active, connections with oscillators in resting state will be weakened. One situation where this occurs is when one oscillator is active whereas the other is completely silent, that is, unstimulated. However, if the oscillators are active in anti-phase, the condition for Hebbian weakening of links is also satisfied. The transition of stimulus-locked activity, combined with anti-phase Hebbian learning, means segmentation can occur because of timing differences in the input.

In this research, it was shown how both segmentation and (re) binding of patterns is achieved in an unsupervised manner by the associative Terman-Wang oscillator network. A range of behaviours was demonstrated, and it was shown that the model is able to display these without altering the parameters governing its dynamics. It must however be noted that the situations Lourenco et al. chose for their experiments, which were copied to a large extent in this research, were chosen for illustrative purposes rather than as a test case for realistic behaviour. In particular, the parameters chosen favour weight alteration on short time scales.

It should also be noted that the external input applied to the oscillators in this model is simulated as a static signal. A more realistic input could be considered in future investigation, using temporally variant signals, such as spiking activity.

4.5 Biological plausibility

If this present model of the AS is in agreement with the way timbre recognition occurs, then what neural structures are required to fulfil all constituent processing tasks?

Meddis and Hewitt (1991) evaluated eight models for sound processing by the IHC's. Although neither model fully replicated physiological results found in literature, the Meddis model was in rather good agreement. Instead of a realistic set of 3500 IHC's (frequency channels), the present model incorporates 64 channels to limit the number of calculations. One drawback of simulating fewer channels is that frequency overlap between adjacent channels is bigger. This does however not preclude the biological plausibility of the model; at worst it causes some loss of the original data. In fact, enhancing the number of frequency channels should only lead to more accurate output patterns.

In the present model spike outputs were directly fed into the delay loops, whereas in the AS spikes are transferred by the cochlear nerve to the olivary complex and cochlear nuclei. A specific cell type in the dorsal cochlear nucleus, the pyramidal build-up cell, might be involved in periodicity filtering as performed here. Build-up cells display a build-up of neural activity when the AS is presented with sound stimuli. This could indicate amplification of periodicities. Furthermore, Frisina et al. (1994) found that in some build-up cells in the dorsal cochlear nucleus signal coding was preserved or even enhanced in the presence of noise.

It is unlikely that single cells are capable of delay loop filtering, since it requires a memory of at least a whole period to integrate periodicities. It is more plausible that neural circuits of cells perform this task. Two cells with recurrent connections may suffice for delay loop filtering, with the first cell as a coincidence detector, and the second one to delay the signal and send it back (figure 19). A neural population with a stochastic distribution of delay loop circuits then can account for filtering a wide range of periodicities.
Inter-spike intervals in biological neurons vary at a millisecond scale, contrary to the neurons in the present model, which display minimum spike intervals of 0.06 ms. The latter is due to the need of a high sampling rate to maintain enough wave information. The AS operates on continuous input and demands less short spike intervals. It was found that phase locking to stimuli of auditory neurons is limited to frequencies below 5000 Hz (Hanekom and Kruger, 2001), which corresponds to inter-spike interval of 0.2 ms. Note that since individual neurons cannot phase-lock to such intervals neural populations are responsible for that.

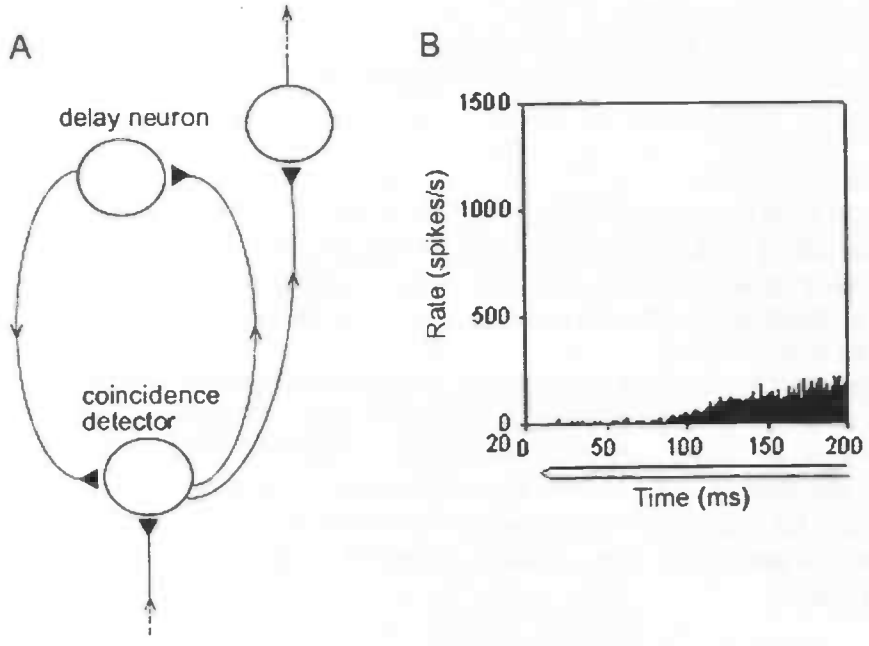


Figure 19. Left: Two neurons constitute a delay loop; Right: At typical discharge pattern of a build-up neuron.

The next stage in audio processing in the AS occurs in the inferior colliculus. Interestingly, this structure has been associated with analysis of periodicities. Langner et al. (1988) found that neurons in the IC responded to amplitude modulated frequencies[2], with oscillation intervals that commonly were multiples of 0.4 ms (i.e. 0.8, 1.2, 1.6 and so on). Periodicity integration by Licklider delay lines could account for these findings.

In the present model autocorrelation output was integrated for each interval channel and after lateral inhibition, it was normalized to a static pattern of standard length. In the cat auditory cortex long-duration-selective neurons have been observed, which only fired in response to noise stimuli when stimulated long enough (He et al., 1997). In that research, some neurons were tuned to a single duration, whereas others fired after a duration threshold had been reached. It is therefore plausible that these neurons have a role as temporal integrators, which transform a time code into place code.

Normalization in the AS is not only a fundamental mechanism, for timbre recognition, but also for speech recognition. Although the exact mechanism behind normalization remains unclear temporal integration probably underlies this phenomenon. In this mode normalization was implemented as a final station of auditory pre-processing, and thus after the patterns were made static, since normalization of temporal information requires accuracy at the microsecond timescale. There is no evidence that such integration accuracy exists.

Lateral inhibition could simply be due to modulation by inhibitory neurons in the cortex, which is a common phenomenon in the brain. Yet, it can also be explained in terms of selective attention. The thalamus, which is important for selective attention, projects inhibitory neurons to the cortex. Thus, when an animal focuses on a sound object, inhibitory thalamic neurons may become active to suppress unexpected audio periodicities.

Conclusion

The associative form of the Wang-Terman oscillator network appears to hold some promising unsupervised learning properties, like time-dependent segmentation and rebinding. Considerations for future research might include a more sophisticated learning rule, as well as investigating realistic temporally structured input.
For better evaluation of its applicability to sound recognition based on timbre, the pre-processing stage should be altered to reduce pattern overlap. Furthermore, although periodicity extraction proved too inaccurate to allow robust distinction of instruments, the segregation network demonstrated closer HD's for intuitively closer related instruments. For a true reliable estimation of periodicity classification, however, larger instrument sets need to be investigated.

[2] Amplitude modulated frequencies are frequency signals coded within a higher carrier frequency, by means of periodic amplitude shifts of the latter.

List of references:

1. Ahissar, E., Abeles, M., Ahissar, M., Haidarliu, S., Vaadia, E. (1998). Hebbian-like functional plasticity in the auditory cortex of the behaving monkey. *Neuropharmacology* vol. 37, pp. 633–55.
2. Cairns, D.E., Baddeley, R.J., and Smith, L.S. (1993). Constraints on synchronizing oscillator networks. *Neural Comp.*, vol. 5, pp. 260-6.
3. Cariani, P.A. (2001). Neural timing nets. *Neural Netw.* vol. 14(6-7), pp. 737-53.
4. Carlson, N.R. (1998). *Physiology of Behavior*, Allyn & Bacon.
5. Frisina, R.D., Walton, J.P., Karcich, K.J. (1994). Dorsal cochlear nucleus single neurons can enhance temporal processing capabilities in background noise. *Exp. Brain. Res.* vol. 102(1), pp.160-4.
6. Gray, C.M., Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc. Natl. Acad. Sci. U S A.* vol. 86(5), pp.1698-702.
7. He J., Hashikawa T., Ojima, H., Kinouchi, Y. (1997). Temporal integration and duration tuning in the dorsal zone of cat auditory cortex. *J. Neurosci.* vol. 17(7), pp. 2615-25.
8. Helmholtz, H. (1954) *On the Sensations of Tone as a Physiological Basis for the Theory of Music* (A.J. Ellis, Trans.): Dover.
9. Izhikevich, E.M. (1999). Weakly Pulse-Coupled Oscillators, FM Interactions, Synchronization, and Oscillatory Associative Memory. *IEEE Transactions On Neural Networks*, vol. 10, pp. 508-26.
10. Krumhansl, C. L., (1989). Why is musical timbre so hard to understand? In S. Nielsen & O. Olsson (Eds.). *Structure and Perception of Electroacoustic Sound and Music* pp. 43-53. Amsterdam: Elsevier.
11. Langner, G., Schreiner, C.E., (1988). Periodicity coding in the inferior colliculus of the cat. Topographical organization, *J. Neurophysiol.* vol. 60(6), pp. 1823-40.
12. Hewitt, M.J., Meddis, R. (1991). An evaluation of eight computer models of mammalian inner hair-cell function. *J. Acoust. Soc. Am.* vol. 90 pp. 904-17.
13. Licklider, J.C.R. (1951). A duplex theory of pitch perception. *Experienta*, vol. 4, pp. 128-34.
14. Lourenco, C., Babloyantz, A., Hougardy, M. (2000). Pattern segmentation in a binary/analog world: unsupervised learning versus memory storing.*Neural. Netw.*. vol. 13(1), pp. 71-89.
15. Marshall, J.A. (1995). Adaptive Perceptual Pattern Recognition by Self-Organizing Neural Networks: Context, Uncertainty, Multiplicity, and Scale. *Neural. Netw.* vol. 8, pp. 335-62.
16. Narayan, S.S., Temchin, A.N., Recio, A.,Ruggero, M.A. (1998). Frequency Tuning of Basilar Membrane and Auditory Nerve. *Science* vol. 282(5395), pp. 1882-4.
17. Schulze, H., Langner, G. (1997). Periodicity coding in the primary auditory cortex of the Mongolian gerbil (Meriones unguiculatus): two different coding strategies for pitch and rhythm? *J. Comp Physiol.* [A] vol. 181(6), pp. 651-63.
18. Shamma, S. (1987). Neural networks for speech processing and recognition. In Proceedings of the International Conferece on *Neural Netw.* pages IV:397--405. IEEE.
19. Sukov, W., Barth, D.S. (2001). Cellular Mechanisms of Thalamically Evoked Gamma Oscillations in Auditory Cortex, *J. Neurophysiol.* Vol. 85(3), pp. 1235-45.
20. Wang, D.L., Brown, G.J. (1999). Separation of Speech from Interfering Sounds Based on Oscillatory Correlation. IEEE *Transactions on neural networks*, vol. 10, pp. 684-97.
21. Wang, D., Buhmann, J., von der Malsburg, C. (1990) Pattern segmentation in associative memory. *Neural Computation* vol. 2, pp. 94-106.