The Detection of Rhythmic Repetition Using a Self-Organising Neural Network

Simon Roberts and Mike Greenhough
Department of Physics and Astronomy
University of Wales College of Cardiff
spxscr@thor.cf.ac.uk

Abstract

An artificial neural network known as SONNET [Nigrin, 1993], which is capable of classifying temporal patterns from a continuous sequential input, is described. When the network is exposed to a sequence of inter-onset-intervals, it is able to detect rhythmic repetition without being supplied with any additional information, e.g. metrical structure. Therefore, the network can be incorporated within a model of rhythm perception to assist with the determination of grouping and metrical structures. Results of network simulations are presented.

1 Introduction

Repetition is an important factor when considering rhythm perception, as it contributes towards the formation of the grouping and metrical structures. When discussing grouping principles, Deutsch [1986] states that "repetition of a subsequence within a sequence induces the listener to group the elements of the subsequence together". Repetition is also a perceptual cue for metre [Palmer and Krumhansl, 1990], as a repeated rhythmic pattern is likely to occur in the same metric position. Lerdahl and Jackendoff [1983] refer to similar passages of music as being "parallel". Parallelism is included in their "Preference Rules" for grouping and metrical structure. Longuet-Higgins and Lee [1982] have developed a model for the perception of musical rhythms. They concluded that the inability of the model to take account of rhythmic repetition was a serious limitation.

The current paper describes an artificial neural network which is capable of detecting rhythmic repetition. Similar research has been carried out by Rosenthal [1989], who developed a computer model that constructs a hierarchical description of a rhythm. Rosenthal's model uses information about metrical structure to assist with the segmentation of the incoming events. The motivation for the work presented in the current paper, is to develop a system which will support the determination of a rhythm's grouping and metrical structures. Therefore, segmentation is based only on the regularities inherent within the input environment. The neural network encodes information about the structure of recurring rhythmic patterns, and is also able to generate expectations. This is beneficial for a model of rhythm perception [Desain, 1992].

2 SONNET Overview

SONNET (Self-Organising Neural Network) [Nigrin, 1993] is an artificial neural network which is capable of classifying patterns from a continually changing input. When the network is exposed to an environment it self-organises, using unsupervised learning, to form stable categories for recurring patterns which exist within the environment. The network creates its own segmentations in response to a stream of incoming events, and is therefore potentially suitable for real-time operation.

SONNET can recognise patterns which are surrounded by extraneous information (embedded patterns) in a context-sensitive manner. That is, when a long, previously learned pattern is presented to the network, SONNET allows the category for that pattern to mask out categories which represent constituent parts of it. (Neural networks which possess this property are known as *masking fields*.) Alternatively, if a sub-part of the long pattern is presented, then a corresponding smaller category is able to classify the shorter pattern. In addition, SONNET allows multiple existing categories to represent novel large patterns.

SONNET also has a number of other desirable features, such as the ability to operate using different learning speeds, create arbitrarily coarse classifications, generate expectations, and represent hierarchical structures.

3 Description of SONNET

The network consists of two fields of cells: F_1 and F_2. The input is applied at F_1, which acts as a short-term memory (STM) and converts a temporal sequence of events into a spatial pattern. The activities of the cells in F_1 are fed forward to each of the cells in the F_2 field via bottom-up excitatory connections. The F_2 cells classify the F_1 patterns. When a novel pattern exists at F_1, many F_2 cells will obtain low activations. After learning, a single F_2 cell will activate strongly whenever its corresponding pattern exists at F_1. The F_2 cell is said to be *committed* to the pattern. The long-term memory representation is stored by the magnitude of the excitatory weights on the bottom-up connections to each F_2 cell. The network architecture described above is based on the Adaptive Resonance Theory

circuits developed by Carpenter and Grossberg. (See for example [Carpenter and Grossberg, 1987]).

As learning progresses, the network self-organises into a masking field, where the F_2 cells have different "sizes". The "size" of an F_2 cell increases with the number of strong bottom-up connections associated with that cell, thus "larger" F_2 cells classify longer patterns. The F_2 cells compete to gain a high activation via lateral, non-uniform inhibitory connections. The inhibitory connectivity pattern is initially uniform, but as the network is exposed to an environment the inhibitory weights self-organise. Eventually, only F_2 cells which respond to overlapping patterns provide mutual inhibition, i.e. patterns which have items in common. Top-down feedback connections from the F_2 field to the F_1 field allow expectation to be introduced. The feedback weights self-organise so that they become approximately parallel to the bottom-up weights.

4 Application of SONNET to the Detection of Rhythmic Repetition

The network was implemented with each F_1 cell representing a particular inter-onset-interval (IOI), measured in musical time, i.e. beats at some metrical level. An additional system is required to identify and track the beats of a particular metrical level, e.g. tactus, to allow the network to be tempo invariant. Also, some form of pre-processing is required to convert an IOI to a place on a spatial map, to allow the correct F_1 cell to be fired. A supervised neural network, such as a multi-layer perceptron trained using back-propagation of error, could be used in the input stage. In the network simulations, the correct mapping was contrived and the F_1 cells were fired accordingly.

After firing, an F_1 cell's activation increases. When the next IOI is detected, its corresponding F_1 cell will fire and the currently active F_1 cells will increase their activity. The activities increase with time to enable expectations to be correctly generated. (See Nigrin [1993] for further details.) However, in the simulations, the top-down feedback connections were disabled, so no expectations were generated. To prevent F_1 overload, a number of the most active F_1 cells are reset after the total activity in the F_1 field has exceeded some threshold. In the simulations, the threshold was set so that F_1 reset would occur after 7 IOIs were held in the STM. This value is the typical STM depth for humans [Miller, 1956]. After an F_2 cell has become committed to a pattern, it is able to *chunk-out* its pattern from the STM, thus reducing the total activity in the F_1 field.

A restriction is placed on the length of a pattern which can be learnt by an F_2 cell. The maximum length was chosen to be commensurate with the number of items which humans can recall before order information becomes confused. This is known as the *transient memory span* and has a typical value of 4 items [Miller, 1956].

Rhythms often only consist of a few different IOIs, so it is necessary for multiple F_1 cells to correspond to the same IOI, to allow repeated IOIs to be present in the STM. So, the number of required F_1 cells is dependent on the number of different IOIs to be represented, and the maximum number of occurrences of a particular IOI to be held in the STM. The latter was taken to equal the STM depth, i.e. 7 F_1 cells relate to the same IOI. Now, when an IOI is presented to the network, one of its associated inactive F_1 cells will fire. The network architecture is shown in Figure 1.

Allowing multiple F_1 cells to represent a single IOI increases the complexity of the network. The reason for this is best explained using an example. Let Q_1, Q_2,...Q_7 denote 7 F_1 cells, each of which represents a quarter-note IOI. If 3 of these cells are activated in the order Q_1, Q_2, Q_3 and an F_2 cell starts to learn this pattern, then this F_2 cell should respond to ♩♩♩ regardless of how it is stored in the STM. Therefore, the F_2 cell should respond equally well to the patterns $Q_3Q_2Q_1$, $Q_5Q_6Q_7$ or any other permutation of 3 F_1 cells representing consecutive quarter-notes.

Nigrin [1993] achieves this behaviour by allowing multiple links to exist from each F_1 cell to each F_2 cell. Each link represents an occurrence of the F_1 cell's associated IOI, in a particular position in the pattern encoded by the F_2 cell. The earlier the position in the pattern, the larger the excitatory weight on the link. This mechanism can self-organise assuming that an F_2 cell can identify which F_1 cells represent the same IOI.

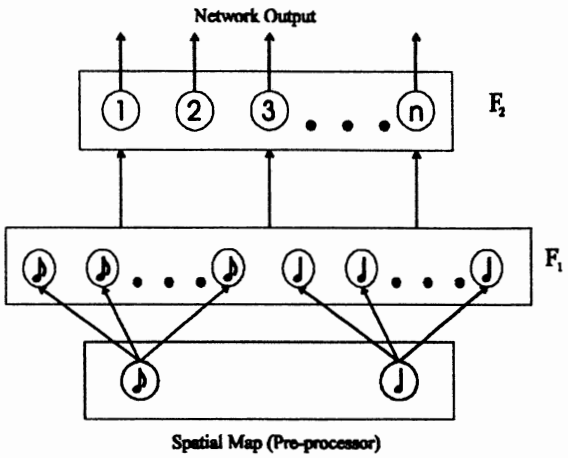

Figure 1: SONNET architecture with multiple F_1 cells representing the same IOI. F_1 is fully connected to F_2 via bottom-up excitatory links, with multiple links from each F_1 cell to each F_2 cell. The lateral inhibitory connections in the F_2 field and the top-down feedback connections are not shown.

We developed slight modifications to enable embedded patterns to be recognised, when multiple occurrences of an IOI could be stored in the STM. For example, suppose an F_2 cell has learned the pattern . The modifications were necessary to allow this cell to recognise that its pattern had recently occurred, when sequences like were held in the STM.

5 Simulations

The same network architecture and network parameters were used for each of the sequences that were presented. The sequences collectively contained 8 different IOIs, so 56 F_1 cells were used to allow 7 occurrences of each IOI to be held in the STM. Twenty-five F_2 cells were used, i.e. a maximum of 25 patterns could be encoded by the network. The cell activities and all of the weights were only permitted to change during a fixed time period after each IOI was presented. This time period will be referred to as the *attention-span*. An attention-span of 0.2s was chosen by considering the shortest IOI, from each sequence, at a typical tempo. (NB: The simulations were run in pseudo real-time.)

A parameter which greatly affects SONNET's ability to detect rhythmic repetition is known as the *learning rate*. A high learning rate enables the network to learn patterns very quickly, but may prevent regularities in the input from being identified. A low learning rate allows recurring patterns and embedded patterns to be learnt, but more sequence presentations are required. A low learning rate was chosen to allow the network to learn regularities from the sequences.

Each sequence was presented 30 times in a continuous manner, i.e. the first IOI of the sequence followed directly on from the last IOI of the previous presentation. The input sequences were based on the rhythms displayed in Figure 2. These rhythms differ by length, complexity of rhythmic structure, and time signature. Figure 3 shows how the network segmented each sequence on the 30th presentation.

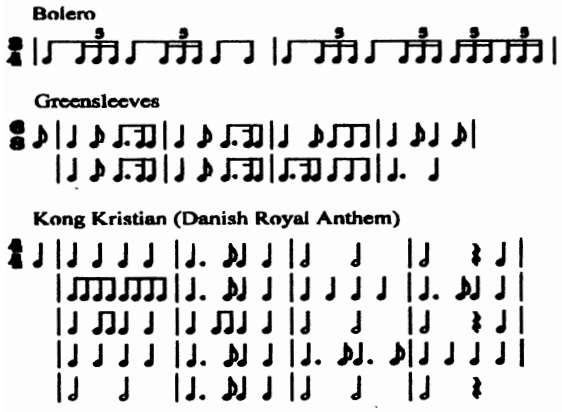

Figure 2: Rhythms used to form input sequences

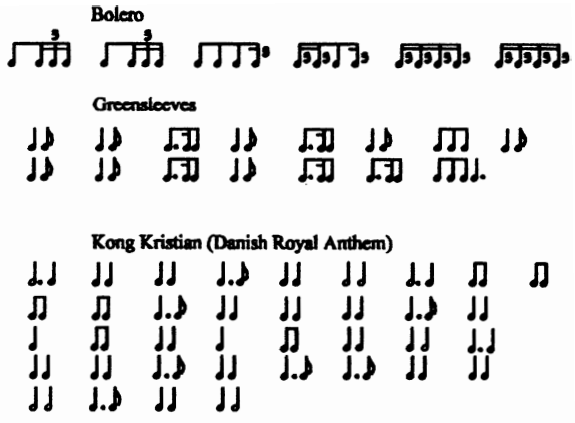

Figure 3: The segmentation of each sequence, performed by the network on the last sequence presentation.

When SONNET was exposed to a sequence, it gradually formed categories, with the most frequently occurring patterns being encoded the earliest. The ability to recognise a recurring pattern is dependent on the number of contexts in which that pattern appears. For example, when the Bolero sequence was presented, the network failed to recognise every occurrence of , because SONNET could not form a category for (which occurs at the end of the first bar), as this pattern always occurred in the same context. Unless short patterns are presented in multiple contexts, SONNET lumps them together with the surrounding IOIs. This problem could be overcome if top-down feedback were used. A committed F_2 cell would then be able to suppress other F_2 cells when its pattern is only partly present at F_1, and the remainder of its pattern is expected.

For Bolero, no F_2 cell encoded the pattern because, after an F_2 cell had become committed to , an uncommitted F_2 cell could only respond to the triplet alone when it was not preceded by an eighth-note (otherwise it received a large inhibitory signal). This only happens at the end of the second bar, so an uncommitted F_2 cell could only respond to the triplet in one context. Consequently, a maximal length pattern of 4 IOIs was formed.

After SONNET had classified the commonly recurring patterns, categories continued to form until a stable representation was obtained for the entire sequence. The number of presentations necessary to achieve a stable representation increased with the complexity of the sequence structure. Bolero required 6 presentations, Greensleeves required 13 and Kong Kristian required 20 presentations.

For Greensleeves, SONNET only failed to detect the repetition of . This was because was always preceded by this pattern and so the network lumped all of these IOIs together.

As Kong Kristian has a more complex structure, the repeating patterns occur in multiple contexts, and therefore SONNET was able to detect all of the

rhythmic repetition. During the presentation of this sequence, SONNET allowed 2 F_2 cells to simultaneously classify their patterns, as there was no overlap between these patterns. This occurred when the F_2 cell encoding a quarter-note combined with the F_2 cell encoding ♩♫ to represent ♩ ♫.

After SONNET was exposed to the Greensleeves or Kong Kristian sequences, some committed F_2 cells had become redundant, i.e. these cells never classified their patterns on later presentations. The reason for this was that the patterns which these cells encoded became partly chunked-out of the F_1 field, as further regularities were classified by other F_2 cells. For example, during the first presentation of Greensleeves, an F_2 cell became committed to ♫♫. After the patterns ♫ and ♩♫ had been encoded by the network, this cell became redundant.

As SONNET only uses regularities in the input to form its categories, the resulting segmentations are not necessarily human-like, as humans involve many principles when grouping IOIs together. If the attention-span increased with IOI, then segmented patterns are likely to end with a long IOI, thus the network would produce more human-like segmentations. This grouping organisation is known as the *gap principle* [Deutsch, 1986]. Music psychologists can benefit from SONNET, because factors which affect the segmentation of a continuous stream of IOIs can be investigated in isolation. For example, the attention-span can be varied to analyse to what extent longer IOIs affect grouping.

6 Further Work

The above simulations served as a preliminary investigation into the performance of SONNET for the detection of rhythmic repetition. A number of alterations are required to create a more compact and elegant system. In the work discussed above, there are 2 distinct representations for time: a particular F_1 cell represents a specific IOI and the F_1 activity pattern encodes the order information. The STM can be implemented so that the F_1 activity pattern represents both the IOIs and the order in which these occur. This is achieved by continuously modifying the F_1 cell activities at short, regular intervals in time, as opposed to only allowing modification to take place during a fixed time period after an event occurs. The firing of an F_1 cell would then simply correspond to the occurrence of an event onset, thus the number of F_1 cells would depend only on the required STM depth. Fewer F_1 cells is advantageous for the simulations, because the computation time increases dramatically with the number of cells in the network. Also, with this STM implementation, no pre-processing is required to convert an IOI to a place on a spatial map. The absence of a pre-processing system overcomes the need to deal with expressive timing in the input stage. Expressive timing can now be processed directly by the SONNET network, as a vigilance parameter controls the coarseness of the classifications.

Multiple SONNET networks can be lumped together to form a hierarchical structure. This property is desirable for the classification of patterns which are inherently hierarchical, such as musical rhythms. Future work will investigate the performance of lumped SONNET networks.

7 Summary

An artificial neural network, known as SONNET, which can classify temporal patterns from a continuous sequential input, has been described. SONNET's ability to detect rhythmic repetition has been demonstrated, by exposing it to 3 different sequences of IOIs. The network is useful for rhythm perception models because grouping structure and metrical structure are dependent on repetition.

References

[Carpenter and Grossberg, 1987] Gail Carpenter and Stephen Grossberg. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, **37**, pp.54-115, 1987.

[Desain, 1992] Peter Desain. A (de)composable theory of rhythm perception. *Music Perception*, **9**(4): pp.439-454, 1992.

[Deutsch, 1986] Diana Deutsch. Auditory pattern recognition. In K. R. Boff, L. Kaufman and J. P. Thomas (Eds.) *Handbook of Perception and Human Performance, Volume 2: Cognitive Processes and Performance*, John Wiley and Sons, New York, 1986.

[Lerdahl and Jackendoff, 1983] Fred Lerdahl and Ray Jackendoff. *A Generative Theory of Tonal Music*, The MIT Press, ISBN 0-262-62049-9, 1983.

[Longuet-Higgins and Lee, 1982] H. C. Longuet-Higgins and Christopher Lee. The perception of musical rhythms. *Perception* **11**, pp.115-128, 1982.

[Miller, 1956] George A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *The Psychological Review*, **63**(2), pp.81-97, 1956.

[Nigrin, 1993] Albert Nigrin. *Neural Networks for Pattern Recognition*, The MIT Press, ISBN 0-262-14054-3, 1993.

[Palmer and Krumhansl, 1990] Caroline Palmer and Carol Krumhansl. Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, **16**(4), pp.728-741, 1990.

[Rosenthal, 1989] David Rosenthal. A model of the process of listening to simple rhythms. *Music Perception*, **6**(3), pp.315-328, 1989.