AISB'01 Convention

21st- 24th March 2001
University of York

Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science

Published by

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour

http://www.aisb.org.uk

ISBN 1 902956 18 9

Printed at the University of York, Heslington, York, Y010 5DD, England

Contents

Towards detection of perceptually similar sounds: investigating self-organizing maps

Christian Spevak; Richard Polfreman; Martin Loomes
Faculty of Engineering and Information Sciences
University of Hertfordshire
College Lane, Hatfield, AL10 9AB
{c.spevak; r.p.polfreman; m.j.loomes}@herts.ac.uk

Abstract

This paper outlines a system for the detection of perceptually similar sounds ('sound spotting'), reports on a series of preliminary experiments and discusses their results. The sound spotting system pursues a frame-based approach and consists of three main stages: an auditory model, a self-organizing map and a pattern matching algorithm. The experiments described examine how different types of self-organizing maps classify a set of test sounds preprocessed by an auditory model and evaluate their performance by means of visualizations and quality measures. With these outcomes in mind we suggest directions for the further development of the sound spotting system.

1 Introduction

Our research addresses a particular problem within the field of content-based retrieval, which can be described as *sound spotting*: the detection of perceptually similar sounds in a given sound document, using a *query by example*, i.e. selecting a prototype sound and searching for occurrences of *similar* sounds. Solutions to this problem would be applicable to indexing/retrieval of sounds in digital archives as well as transcription and analysis of non-notated music.

Over the last ten years a number of researchers have investigated connectionist approaches to model the perception of timbre (Feiten and Gunzel, 1994, Toiviainen, 1997; Toiviainen et al., 1998; Cosi et al., 1994; De Poli and Prandoni, 1997). Sounds are preprocessed with a simplified model of the auditory periphery, and the resulting feature vectors are classified by means of a self-organizing map, which projects multidimensional input vectors onto a low-dimensional topological surface. An introduction to this area including a brief literature survey has recently been given by Toiviainen (2000).

Our concept attempts to extend these models by dealing with evolutions of timbre, pitch and loudness in a dynamic, frame-based approach involving the stages listed below.

The raw audio data is preprocessed with an *auditory model* to obtain a perceptually relevant representation; for the purpose of data reduction the signal is subsequently divided into short frames, each of them consisting of a feature vector.

A *self-organizing map* (SOM) is employed to perform a vector quantization and a topology-preserving mapping of the feature vectors. At this stage a sound signal corresponds to a trajectory on the map.

Finally pattern matching is applied to detect trajectories or sequences of feature vectors 'similar' to a selected prototype. We are currently testing a *Dynamic Programming* algorithm (DP matching).

This paper evaluates the performance of different self-organizing maps—varying in size, dimensionality, type of lattice, and shape—in combination with an auditory model and a set of test sounds. The results of these experiments lead to further suggestions concerning the structure of the proposed sound spotting system.

Experiments investigating the effect of different auditory representations combined with one particular type of SOM have already been discussed in a previous paper[1] (Spevak and Polfreman, 2000).

2 System components

2.1 Auditory model

The auditory model used here combines an auditory filterbank and an inner hair cell model. The filterbank consists of fourth order gammatone filters, which provide a good fit to human auditory filter shapes (Patterson and Holdsworth, 1996). The inner hair cell model, developed by Meddis (1986), simulates mechanical to neural transduction in each filter channel by modeling the transmitter

[1] The SOMs consisted of approximately 80 units, arranged in a hexagonal, sheet-shaped lattice. The auditory representations examined included the gammatone filterbank in combination with an inner hair cell model, Lyon's cochlear model, and mel-frequency cepstral coefficients (MFCC). The gammatone model produced the most convincing results and was therefore chosen for this study.

release from hair cells into the synaptic cleft Its output represents the instantaneous spike probability in a post-synaptic auditory nerve fiber, showing features such as adaptation and phase locking to low-frequency periodic stimuli

The experiments were carried out with 64 filter channels covering a frequency range from 100 Hz to 10 kHz, using a sampling rate of 22 05 kHz To reduce the amount of data[2], but still be able to track quick changes of pitch or timbre, the output was lowpass filtered and decimated to a frame rate of 100 Hz

2.2 Self-organizing map

Self-organizing maps constitute a particular class of artificial neural networks, developed by Kohonen (1997) and inspired by brain maps, such as the tonotopic map of pitch in the auditory cortex A SOM is able to map high-dimensional input signals onto a low-dimensional grid while preserving the most important topological relations, so that similar input signals are usually located close to one another The self-organization takes place during an unsupervised training phase the preprocessed data is repeatedly presented to the network, which adapts its weight vectors according to the topology of the input signals, thus forming a feature map

2.2.1 The SOM algorithm

In the following the basic SOM algorithm, also known as *incremental learning*[3], is briefly described A SOM consists of neurons arranged on a low-dimensional lattice Each neuron is associated with an n-dimensional weight vector $\mathbf{m} = [m_1, m_2 \quad m_n]$, where n corresponds to the dimension of the input signal The weight vectors are initialized randomly or linearly according to the distribution of the training data Training is performed iteratively, in each step, a sample vector $\mathbf{x}$ is chosen randomly from the set of input data, and the distance to each of the weight vectors is calculated The neuron whose weight vector $\mathbf{m}_c$ is most similar to the input vector $\mathbf{x}$, as defined by the condition

$$\|\mathbf{x}(t) - \mathbf{m}_c(t)\| = \min_i \|\mathbf{x}(t) - \mathbf{m}_i(t)\|, \qquad (1)$$

is identified as the *best-matching unit* (BMU) or the *winner* ('winner-takes-all' function) Subsequently the weight vectors of the best-matching unit and its topological neighbours are updated toward the input vector The SOM update rule is expressed by the following equation

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + \alpha(t)h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)], \qquad (2)$$

[2]The auditory model initially produces a 64-element vector for each sample i e it increases the amount of data significantly

[3]We actually used a faster variant the *batch training algorithm* which is functionally equivalent

Table 1 Sound set comprising simple synthesized tones and noise signals '<' denotes increasing and '>' decreasing amplitude

#	*Waveform, frequency*	#	*Waveform, frequency*
01	noise band, 0–1 kHz	12	sine octaves, 2/4 kHz
02	noise band, 1–5 kHz	13	sine oct , 400/800 Hz
03	white noise	14	sine <, 1 kHz
04	square, 100 Hz	15	sine >, 1 kHz
05	square, 1 kHz	16	sine 100 Hz
06	square < 1 kHz	17	sine, 1 kHz
07	square >, 1 kHz	18	sine 500 Hz
08	square, 500 Hz	19	sine 5 kHz
09	square, 5 kHz	20	triangle, 1 kHz
10	sine sweep 0–10 kHz	21	triangle, 100 Hz
11	sine and noise bursts	22	triangle 500 Hz
		23	triangle 5 kHz

where $\mathbf{m}_i$ denotes the weight vector of the i th neuron, $\mathbf{x}$ the input vector, t the discrete time coordinate, α the learning rate, and h_{ci} the neighbourhood kernel around the winner unit c

The training is usually performed in two phases the ordering phase, typically consisting of 1000 steps, and the fine-tuning phase, extending across 10,000 steps or more, depending on the size of the map During the ordering phase both the learning rate and the neighbourhood kernel decrease from their large initial values to small values used for fine-adjustment, e g the neighbourhood radius may shrink from half the diameter of the network to the distance between adjacent neurons

3 Methodology

3.1 Sound set

The test sound set comprised 23 monophonic synthesized signals of 2 s duration, sampled at 22 05 kHz Each sample consists of a 1 s sound event framed by half a second of silence The set includes white and band-limited noise, steady sine, triangle and square wave signals at various frequencies, a sine pitch sweep from 0–10 kHz, sine octaves, sine and square waves with increasing and decreasing amplitude respectively, and a sample of quickly alternating tone and noise bursts Table 1 provides a complete list

3.2 Tools

The experiments have been carried out in *Matlab*®, an integrated environment for numeric computation, visualization and programming The simulation of auditory models and neural networks was facilitated by the use of specialized 'toolboxes' in addition to the main program, in particular the *Auditory Toolbox* (Slaney, 1998) and the *SOM Toolbox for Matlab 5* (Vesanto et al , 2000)

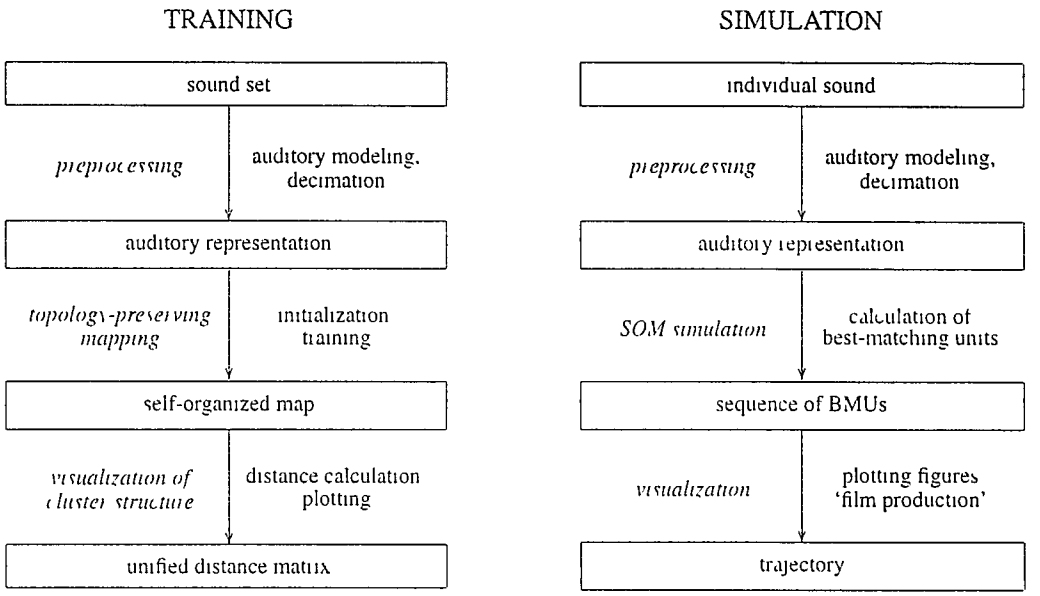

Figure 1 Block diagram showing the individual steps carried out during training and simulation of a self-organizing map with a set of sounds

3.3 Outline of the experiments

Neural network experiments are typically made up of two main parts *training* and *simulation*. In this case the training phase involved the preprocessing of the entire sound set with an auditory model and the decimation to a lower frame rate, the initialization and training of a SOM, and finally the visualization of its cluster structure. The simulation phase served to determine the trajectory of a particular sound by finding the corresponding sequence of best-matching units and producing a visualization. Figure 1 gives an overview of the individual stages and processing steps.

3.3.1 Visualization of the cluster structure

The *U-matrix* or *unified distance matrix* shows the cluster structure of a self-organized map by visualizing the vector space distances between adjacent map units in different shades of grey. Clusters of similar units stand out as light patches, surrounded by darker borders. This representation was used to visually inspect the SOM once the training was completed. (An example is shown in Figure 3.)

3.3.2 SOM quality analysis

Each trained SOM was subjected to a quality analysis by determining the *average quantization error* and the *topographic error*. The former measures the goodness of fit between the training data and the SOM weights. It is defined as the mean of the Euclidean distances $\|\mathbf{x} - \mathbf{m}_c\|$ between the training vectors $\mathbf{x}$ and their respective BMU $\mathbf{m}_c$. The topographic error quantifies the accuracy of the SOM in preserving the topology of the training data. It indicates the percentage of training vectors for which the BMU and the second-BMU are not adjacent map units.

3.3.3 Visualization of trajectories

The sequence of BMUs corresponding to a sound can be visualized as a *trajectory* on the SOM's two-dimensional lattice. To analyze the trajectories corresponding to the tests sounds we developed an animated representation where the trajectory is built up frame by frame in slow motion. The representation includes a waveform picture of the sound with a moving pointer indicating the current position. Figure 2 shows an example of a still frame.

Three-dimensional SOMs were visualized by a 'half-open box', consisting of one vertex and the three adjacent faces seen from the inside, where the position of the current BMU was indicated by a red dot and its projections onto each of the three faces.

4 Results

4.1 SOM size

How do different SOM sizes influence the mapping of a given data set, and what is the *ideal* size? Vesanto et al. (2000) recommend to derive the number of map units from the number of training vectors, using the heuristic rule $n_{\text{MU}} = 5\sqrt{n_{\text{TV}}}$. Following this equation we created a 'medium-sized SOM and compared it to a 'small' and a 'large' map comprising of $\frac{1}{4}$ and 4 times the number of

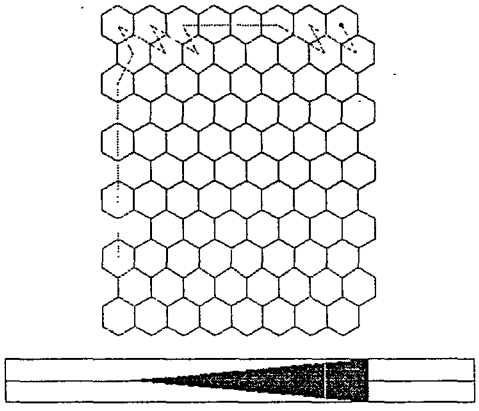

Figure 2: Still frame from a film visualizing the trajectory produced by a square wave with increasing amplitude

Table 2: Comparison of quantization error, topographic error, BMU percentage and training lengths for three different SOMs, consisting of a two-dimensional, hexagonal, sheet-shaped lattice. The training data was derived from approximately 45 s of sound. Durations were recorded on a Pentium III 450 MHz PC.

map size	'small'	'medium'	'large'
number of units	88	340	1353
	(11 × 8)	(20 × 17)	(41 × 33)
quantization error	0.00186	0.00044	0.00023
topographic error	4.2%	4.9%	3.6%
BMU percentage	85%	62%	51%
ordering phase	1 cycle	1 cycle	3 cycles
	(< 1 s)	(5 s)	(115 s)
fine tuning phase	1 cycle	3 cycles	12 cycles
	(1 s)	(15 s)	(562 s)

map units, respectively. Table 2 details the sizes and typical measurements, such as errors and training lengths.

Judging by the quality measures the large SOM shows the best adaptation to the training data: it has the lowest average quantization error as well as the lowest topographic error[4]. However, the training time increases disproportionately to the SOM size, because the larger number of units requires more training cycles as well as more computations during each cycle. Training is performed quite efficiently for the small and medium-sized SOMs and becomes very expensive for the large SOM.

The measure *BMU percentage* expresses the share of the SOM units that serve as a best-matching unit at least once when the complete training set is presented to the ordered map. In this context the measure provides a useful indication of the SOM's efficiency, because the SOM is presented with the whole range of data during the training phase and is not expected to generalize to new data. Therefore interpolated units that do not act as BMUs are largely superfluous. The BMU percentage and thus the efficiency of the SOM clearly increase for smaller maps.

[4]The differences between the topographic errors are not very significant, because there is no obvious correlation between SOM size and topographic error

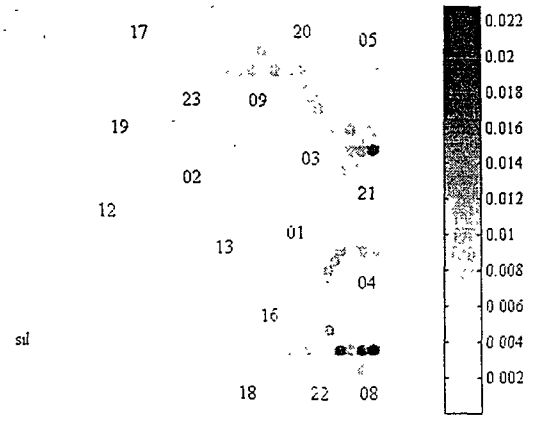

Figure 3: Labeled U-matrix of a 'medium'-sized SOM trained on the preprocessed test sounds. The numbers on the map refer to the steady state locations of the test sounds listed in Table 1. 'Silence' is located near the lower left corner of the map, surrounded by a large cluster of similar weight vectors.

The U-matrices of the larger maps exhibited a distinct cluster structure, as shown in Figure 3, where clusters of neurons with similar weights are separated from one another by larger distances in weight space. Such a cluster structure complicates the targeted detection of similar trajectories, because the pattern recognition system would have to distinguish between units located in the same cluster and units located in different clusters. However, the exact definition of a *cluster* is ambiguous, because the borders often become blurred. Therefore 'small' SOMs that reduce most of the clusters to single units seem to be more appropriate. And on top of that smaller SOMs are computationally much more efficient. But since they have a lower resolution, it is important to define a criterion for the desired minimum resolution. Labeling the best-matching units corresponding to the steady states of the test sounds showed that in this case even the small SOM was able to 'resolve' the different sounds. However, for less repetitive sets of sounds[5] it may be more appropriate to use a 'medium'-sized SOM.

4.2 Dimensionality

It is theoretically possible to construct SOMs that span an arbitrary number of dimensions, but more than three dimensions are very rarely used in practical applications.

[5]Since half of the training data consisted of silence and most sounds had a steady spectrum the actual number of *different* training vectors was much lower than the overall number

Table 3 Comparison of quantization error, topographic error and BMU percentage for one-, two-, and three-dimensional SOMs consisting of approximately 90 units arranged in a rectangular lattice

dimensionality	1	2	3
number of units	88	88	90
	(88)	(11 × 8)	(6×5×3)
quantization error	0.0041	0.0017	0.0021
topographic error	2.4%	16.5%	5.7%
BMU percentage	83%	80%	71%

Two-dimensional maps are most common, because they lend themselves very well to visualization. De Poli and Tonella (1993) classified sounds with three-dimensional maps to construct a three-dimensional *timbre space* originally derived from similarity ratings by Grey (1977).

We studied the performance of one-, two- and three-dimensional SOMs of similar size, using a sheet-shaped rectangular lattice[6]. The exact dimensions and the resulting quality measures are listed in Table 3.

Interestingly the one-dimensional SOM had the lowest topographic error and the highest BMU percentage, indicating an accurate topological organization and high efficiency. On the other hand it showed the highest quantization error, i.e. it approximated the individual training vectors less closely than the higher-dimensional maps. For the two-dimensional SOM the performance seemed to deteriorate in connection with a rectangular lattice, resulting in an exceptionally high topographic error (cf. Section 4.3). The three-dimensional SOM had the lowest BMU percentage and medium error values. Altogether it did not seem to provide any clear advantages over the two-dimensional map.

4.3 Lattice

Comparing the performance of two dimensional SOMs differing only in their lattice structure revealed a striking discrepancy between the topological error values, as shown in Table 4. The topological error was much higher for a rectangular lattice than for a hexagonal one. This may be determined by the fact that a unit in a hexagonal lattice is surrounded by six equidistant neighbours, while a unit in a rectangular lattice has four next neighbours and four 'diagonal' neighbours. The latter do not count as neighbours in the calculation of the topological error, but they are included by the Gaussian neighbourhood function used to update the weight vector during the training phase. The differences in quantization error and BMU percentage are less noticeable, but altogether the hexagonal lattice seems to be preferable. Kohonen recommends it particularly for visualization, because a rectangular grid tends to favour horizontal and vertical directions (Kohonen 1997, p. 120).

[6] A strictly hexagonal lattice can only be realized in two dimensions.

Table 4 Comparison of quantization error, topographic error and BMU percentage for hexagonal and rectangular lattice SOMs

lattice	hex	rect	hex	rect
units	88	88	340	340
	(11 × 8)	(11 × 8)	(20 × 17)	(20 × 17)
quant. err.	0.00186	0.00168	0.00044	0.00043
topogr. err.	4.2%	16.5%	4.9%	12.5%
BMU %	85%	80%	60%	62%

4.4 Shape

The plane sheet is not the only possible shape for a self-organizing map—it can be 'wrapped around' in one or two dimensions, resulting in a *cylindric* or *toroidal* map, respectively. However, neither of these alternative shapes seemed to be particularly well suited to our data: the error values increased, and the visualizations looked confusing because clusters stretched across the edges. Toroidal maps are only recommended if the data itself has a cyclic structure. Musical keys for instance can be arranged in a *circle of fifths*. Leman (1994) successfully employed toroidal SOMs for tone centre recognition, and Purwins et al. (2000) further developed the system to track modulations in tonal music.

4.5 Summary

Our investigation of self-organizing maps combined with an auditory model to classify sounds suggested that a relatively small SOM based on a hexagonal, sheet-shaped lattice would be the preferable solution. The different sounds were clearly separated on the map and grouped according to their pitch, or fundamental frequency. However, even with the 'optimal' SOM the organization of the sounds on the map was far from perfect when compared to our perception: pairs of sounds having the same distance on the map could be either perceptually similar or entirely different, depending on the respective cluster structure.

5 Discussion

A self-organizing map can be a powerful visualization tool, but it seems to be less suitable to actually quantify 'similarity'. Because of the inhomogeneous distribution of weight vectors, the distance between best-matching units on the map does not constitute a particularly suitable distance measure for the corresponding sounds. Toivainen (1996) corroborates this by stating that correlations between subjective similarity ratings and distance metrics on the SOM were usually lower than those obtained using the distances between the preprocessed feature vectors. He argues that the dimensionality reduction in the SOM distorts the metrical relationships between the input vectors.

Considering these results there are several possibilities to complete the sound spotting system described in the introduction. the pattern matching algorithm could either be applied to the index number of the best-matching units (performing a *string matching* task) or to the corresponding weight vectors, or directly to the feature vectors produced by the auditory model The former two variants reduce the SOM to a *vector quantization* device (neglecting the topology-preserving mapping), while the latter bypasses it completely

Our future research will examine these possibilities in detail and assess their performance by correlating it with similarity ratings obtained from expert listeners, using a more comprehensive set of sounds

References

Pietro Cosi, Giovanni De Poli, and Giampaolo Lauzzana Auditory modelling and self-organizing neural networks for timbre classification *Journal of New Music Research*, 23(1) 71–98, 1994

Giovanni De Poli and Paolo Prandoni Sonological models for timbre characterization *Journal of New Music Research* 26(2) 170–197, 1997

Giovanni De Poli and Paolo Tonella Self-organizing neural network and Grey's timbre space In *Proceedings of the International Computer Music Conference (ICMC)*, pages 260–263, Tokyo, 1993

Bernhard Feiten and Stefan Gunzel Automatic indexing of a sound database using self-organizing neural nets *Computer Music Journal*, 18(3) 53–65, 1994

John M Grey Multidimensional perceptual scaling of musical timbre *Journal of the Acoustical Society of America*, 61(5) 1270–1277, 1977

Teuvo Kohonen *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences* Springer, Berlin 2nd extended edition, 1997. 1st edition 1995

Marc Leman Schema-based tone center recognition of musical signals *Journal of New Music Research*, 23 (2) 169–204, 1994

Ray Meddis Simulation of mechanical to neural transduction in the auditory receptor *Journal of the Acoustical Society of America*, 79(3) 702–711, March 1986.

Roy D Patterson and John Holdsworth A functional model of neural activity patterns and auditory images In William A Ainsworth, editor, *Advances in Speech Hearing and Language Processing*, volume 3 JAI Press London, 1996

Hendrik Purwins, Benjamin Blankertz, and Klaus Obermayer A new method for tracking modulations in tonal music in audio data format In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 270–275, Como, Italy, July 2000

Malcolm Slaney Auditory Toolbox Version 2 Interval Technical Report 1998-010, Interval Research Corporation, Palo Alto, CA, 1998

Christian Spevak and Richard Polfreman Analyzing auditory representations for sound classification with self-organizing neural networks In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx-00)*, pages 119–124, Verona Italy, December 2000 `http://www.sci.univr.it/~dafx/`

Petri Toiviainen Optimizing auditory images and distance metrics for self-organizing timbre maps. *Journal of New Music Research*, 25(1) 1–30, 1996

Petri Toiviainen Optimizing self-organizing timbre maps Two approaches In Marc Leman editor, *Music, Gestalt, and Computing Studies in Cognitive and Systematic Musicology*, pages 337–350 Springer-Verlag, Berlin, Heidelberg, 1997

Petri Toiviainen Symbolic AI versus connectionism in music research In Eduardo Reck Miranda, editor, *Readings in Music and Artificial Intelligence*, volume 20 of *Contemporary Music Studies*, chapter 4, pages 47–67 Harwood Academic Publishers, Amsterdam, 2000

Petri Toiviainen, Mari Tervaniemi, Jukka Louhivuori, Marieke Saher, Minna Huotilainen, and Risto Naatanen Timbre similarity Convergence of neural, behavioral, and computational approaches *Music Perception*, 16(2) 223–242, 1998

Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas SOM Toolbox for Matlab 5 Technical Report A57, Helsinki University of Technology, April 2000