Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy, December 7-9,2000

ANALYZING AUDITORY REPRESENTATIONS FOR SOUND CLASSIFICATION WITH SELF-ORGANIZING NEURAL NETWORKS

Christian Spevak and Richard Polfreman

Music Department
University of Hertfordshire, UK
{c.spevak,r.p.polfreman}@herts.ac.uk

ABSTRACT

Three different auditory representations—Lyon's cochlear model, Patterson's gammatone filterbank combined with Meddis' inner hair cell model, and mel-frequency cepstral coefficients—are analyzed in connection with self-organizing maps to evaluate their suitability for a perceptually justified classification of sounds. The self-organizing maps are trained with a uniform set of test sounds preprocessed by the auditory representations. The structure of the resulting feature maps and the trajectories of the individual sounds are visualized and compared to one another. While MFCC proved to be a very efficient representation, the gammatone model produced the most convincing results.

1. INTRODUCTION

The fundamental problem investigated in this research is 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 'similar' occurrences. A solution to this problem would have applications in the analysis of musical works, transcription of non-notated music, and indexing/retrieval of sounds in self-contained documents in general.

Our research is developing a modular system consisting of the following stages:

- preprocessing of the raw audio data with an auditory model to simulate the auditory pathway and extract perceptually relevant features; subsequent data reduction by dividing the signal into short frames,
- topology-preserving mapping of each frame onto a self-organizing map (SOM),
- detection of similar trajectories on the map by means of sequence comparison.

The third stage has not yet been implemented and will be the subject of future research.

Over the last ten years several studies have successfully classified timbre by means of artificial neural networks, using auditory models and self-organizing maps. Results for limited sets of sounds have been published by Feiten and Günzel [1], Toiviainen et al. [2, 3] and Cosi, De Poli et al. [4, 5]. Since the structure of the auditory model, the topology and size of the self-organizing maps, and the set of test sounds are different for each approach, the results are difficult to compare. In addition, our research aims at dealing with timbre evolutions rather than classifying steady state samples.

This paper describes an evaluation of the performance of three different auditory representations in combination with a two-dimensional SOM and a set of 23 test sounds, covering a wide range of timbre, pitch, and amplitude values. The analyzed auditory representations are Lyon's passive cochlear model, a gammatone filter bank combined with Meddis' inner hair cell model, and mel-frequency cepstral coefficients.

2. AUDITORY REPRESENTATIONS

2.1. Lyon's cochlear model

The passive cochlear model described by Lyon [6] and Slaney [7] transforms the sound signal into a probability of firing along the auditory nerve, using the following components: a preemphasis filter to simulate the frequency response of the middle and outer ear, a broadly tuned cascade of lowpass filters (96 stages at 22 kHz sampling rate) to model the traveling wave on the cochlea, half wave rectifiers to implement the detection nonlinearity of the inner hair cells, and four stages of automatic gain control with different time constants to simulate adaptation and masking.

2.2. Gammatone filterbank and Meddis' IHC model

The second model evaluated consists of an auditory filterbank described by Patterson et al. [8, 9] and implemented by Slaney [10], and an inner hair cell (IHC) model developed by Meddis [11]. The filterbank is based on fourth order gammatone filters, which provide a good fit to human auditory filter shapes (cf. Cooke [12, p. 16]). The experiments were carried out with 64 filter channels covering the frequency range from 100 Hz to 10 kHz.

Meddis' IHC model simulates mechanical to neural transduction in each filter channel by modelling the transmitter release from hair cells into the synaptic cleft. The 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.

2.3. Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients (MFCC), introduced by Davis and Mermelstein [13], constitute a parametric sound representation widely used in automatic speech recognition systems. MFCC has also been successfully applied to timbre analysis [14, 5]. The signal is passed through a mel-spaced filterbank (based on FFTs), converted to a logarithmic scale, and then submitted to a cosine transform. MFCC provide a substantial data reduction, because a few coefficients are sufficient to represent the *cepstrum* of the

acoustic signal. In this case the power-related first coefficient C_0 was discarded, because its large variance would have dominated the organization of the SOM.

3. SELF-ORGANIZING MAPS

Self-organizing maps (SOMs) constitute a particular class of artificial neural networks, developed by Teuvo Kohonen [15] 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.

3.1. The SOM algorithm

In the following the basic SOM algorithm, also known as *incremental learning*, is briefly described. A SOM consists of neurons arranged on a low-dimensional lattice. Each neuron is associated with an n-dimensional weight vector $\mathbf{m} = [m_1, m_2, \ldots, m_n]$, where n corresponds to the dimension of the input signal. First of all the weight vectors are initialized—either randomly or linearly according to the distribution of the training data. The 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-take-all' function). After that, 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:

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

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.

4. METHODOLOGY

4.1. Overview

A neural network experiment usually requires two main processes: *training* and *simulation*. In this case the training phase involved the preprocessing of the complete sound set with one of the auditory models and the decimation to a lower frame rate, the initialization and training of a SOM, and finally a quality and cluster analysis. The simulation phase served to determine the trajectory of a particular sound by finding the corresponding sequence of best-matching units and producing a suitable visualization.

4.2. Tools

The experiments were carried out in MATLAB®, an integrated environment for numeric computation, visualization, and programming. In addition to the main programs we used Slaney's Auditory Toolbox [16] and the SOM Toolbox developed by Vesanto and colleagues [17].

4.3. Sound set

A prerequisite for the analysis of auditory representations and self-organizing maps was a set of well-defined test sounds that met the following requirements: it should cover a range of different timbres as well as different pitch and loudness values, but still contain subsets of sounds with common timbre, pitch, or loudness. The sounds should be short and simple enough to produce a visually clear trajectory on the SOM, but also show some dynamic evolutions in pitch and loudness. A short period of silence at the end of each sound would be useful to trace the decay characteristics of the auditory models.

The actual sound set comprises 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. The complete set is listed in Table 1.

4.4. Calculation of the auditory models

The auditory models were calculated at a sampling rate of 22.05 kHz, allowing the processing of frequencies up to ∼11 kHz. To reduce the amount of data passed on to the SOM, but still be able to track quick changes of pitch or timbre (such as a pitch sweep or a sudden attack), the output frame rate was reduced to 100 Hz. In the first two models this was achieved by lowpass filtering the output and subsequently picking every 100th value, in the MFCC model the frame rate was determined by the step size of the FFT.

4.5. Computation of the SOMs

For each of the auditory representations an individual SOM was trained with the complete set of preprocessed test sounds. The SOMs consisted of approximately 80 units, arranged in a two-dimensional hexagonal grid. These parameters were chosen considering the results of experiments involving different SOM sizes, shapes, and topologies (see Section 5.1). The exact size was determined on account of the ratio between the first two principal components of the training data. The weight vectors were initialized linearly along these components.

Training was performed by the *batch training* algorithm [17], a faster variant of the SOM algorithm described in Section 3.1. Instead of adjusting the weights to each individual data vector, the complete training data is presented to the map before the weight

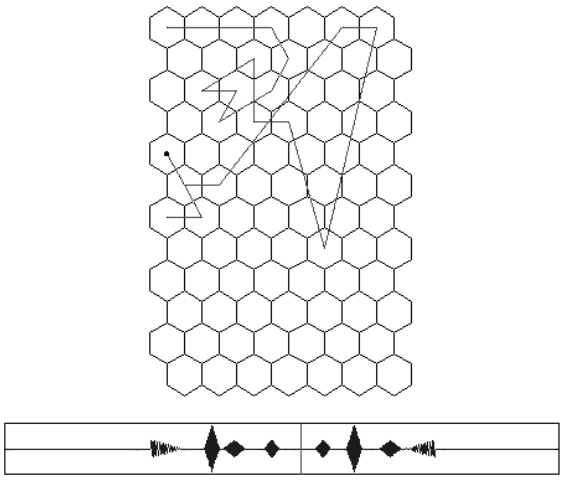

Figure 1: Still frame from a film visualizing the trajectory produced by a sequence of quickly alternating tone and noise bursts, preprocessed with Lyon's cochlear model.

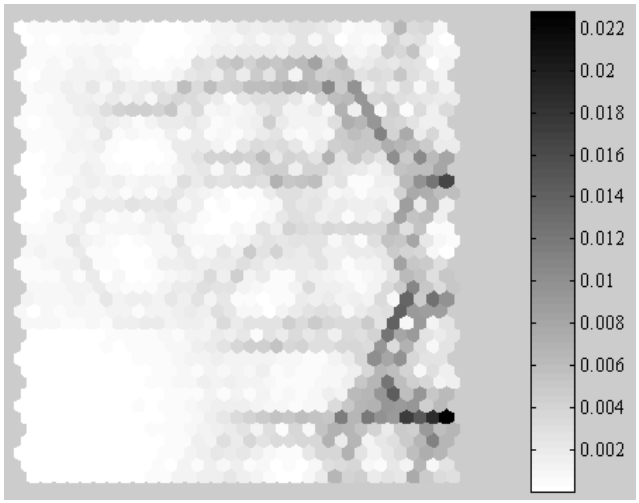

Figure 2: U-matrix of a 20×17 SOM in combined with the gammatone/IHC preprocessing. The shades of grey represent the distances between adjacent units in the weight space. The map units themselves are coloured according to the mean of the distances to all their neighbours. Cluster borders are indicated by darker colours.

vectors are replaced by a weighted average of the data vectors that were in their neighbourhood.

After having completed the training, the SOM was simulated with each individual sound to record the respective sequence of *best-matching units*. A BMU was defined as the closest weight vector using the Euclidean distance measure

$$d_E(\mathbf{x}, \mathbf{m}) = \|\mathbf{x} - \mathbf{m}\| = \sqrt{\sum_i (x_i - m_i)^2}, \qquad (3)$$

where x_i and m_i are the vector components of $\mathbf{x}$ and $\mathbf{y}$.

4.6. Visualization of the output

The sequence of BMUs corresponding to a sound can be visualized as a trajectory on the SOM's two-dimensional lattice. Simply connecting the BMUs by lines causes some problems, though: such a representation does neither show the direction of the trajectory nor the duration of stay at a particular unit. While these problems could be solved by introducing arrowheads and variable marker sizes, it would not be possible to represent a to-and-fro movement between two or more units, which occurred quite often. Therefore we decided to develop an animated representation, where the trajectory is built up frame by frame. The representation includes a waveform picture of the sound with a moving pointer indicating the current position. Figure 1 gives an example of a still frame.

5. RESULTS

5.1. Different SOM sizes

The size of a SOM is usually determined by the amount of training data. A heuristic formula given by Vesanto et al. [17] calculates the number of map units as $n_{\mathrm{MU}} = 5\sqrt{n_{\mathrm{TD}}}$. Applying this formula to our training data resulted in SOMs comprising approximately 340 units. On completing the training these maps showed a distinct cluster structure, i.e. groups of neurons having similar weights were separated from one another by larger distances in weight space. The cluster structure of a SOM can be visualized by means of a unified distance matrix or U-matrix, as shown in Figure 2.

Since such a large cluster structure complicates the targeted detection of similar trajectories[1], we moved to smaller SOMs (approximately one fourth of the size given above), where the clusters are mostly reduced to single units[2]. Apart from that smaller SOMs are computationally much more efficient. The double assignment of some units to different, but similar sounds (cf. Figures 3 and 5) illustrates that the SOM performs a vector quantization whose resolution corresponds to the map size.

5.2. Lyon's cochlear model

The locations of the steady state BMUs on a 12×7 SOM in connection with Lyon's cochlear model are diplayed in Figure 3. The distribution is reproducible, because it is based on a deterministic initialization. Silence, the most frequent 'event' in the whole sound set, is mapped to the upper left corner. This is the place where all sound trajectories start and end. Since the 'silence vector' is presented to the SOM so often during the training phase, it influences a large number of surrounding units, which can be visualized by a U-matrix (cf. Figure 2, lower left corner.). These units serve the trajectories as 'stopovers' during the attack and decay phase: instead of jumping instantly from silence to the steady state locations shown in Figure 3, the trajectories move forward in small steps, often in a zigzag. The delay is caused by the filter used in the decimation process—it is missing in the MFCC-trajectories.

[1]The pattern recognition system would have to distinguish between units located in the same cluster and units located in different clusters instead of regarding all units simply as different states. However, the exact definition of a *cluster* is ambiguous, because the borders often become blurred.

[2]Sound signals that are less redundant than the test sounds used here might still require larger SOMs.

No.	*Waveform, frequency*	*No.*	*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

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

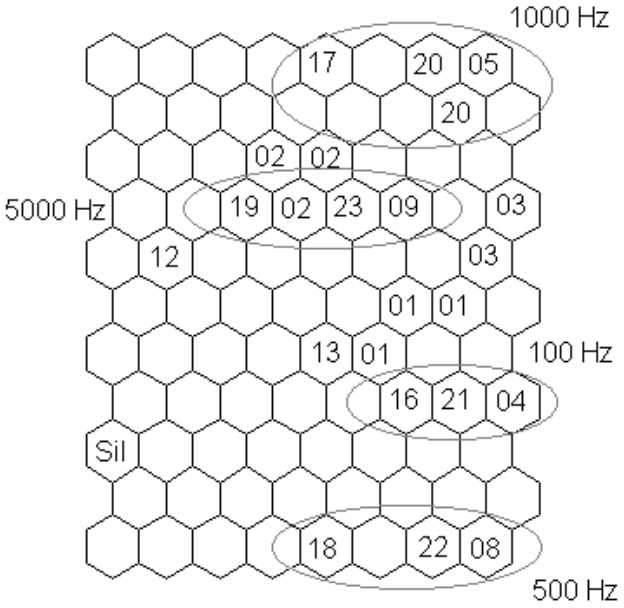

Figure 4: Locations of the steady state BMUs on a 11×8 SOM in connection with the gammatone/IHC model.

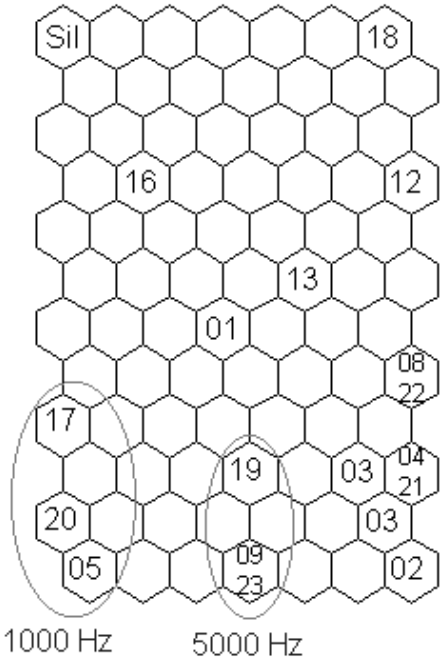

Figure 3: Locations of the steady state BMUs on a 12×7 SOM in connection with Lyon's cochlear model. The numbers correspond to the sounds in Table 1, 'Sil' stands for 'Silence'.

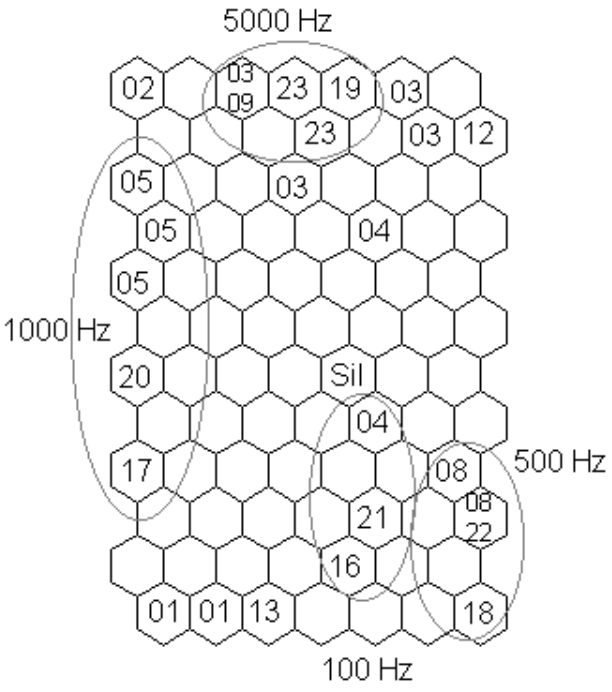

Figure 5: Locations of the steady state BMUs on a 12×7 SOM in connection with the MFCC preprocessing.

When a loud sound stops abruptly, it takes the trajectory approximately 15 frames (150 ms) to return to the point of silence[3]. If a new signal sets in during that time, the trajectory does not return to the origin at all. This applies for instance to the quickly alternating tone and noise bursts, as shown in Figure 1.

The noise signals 02 and 03 are located diametrically opposed to silence, which is interesting, because from a spectral point of view white noise is exactly the opposite of silence. The 1 kHz tones are mapped close to one another, as well as the 5 kHz tones. Square and triangle waves with common fundamental frequency are even mapped to the same unit, which is perceptually understandable. Pitch is thus a crucial factor for the organization of the SOM.

Variations in sound level influence the mapping only if the absolute sound level lies below a certain threshold; a change from medium to higher level is usually not reflected in the trajectory. This can be explained by the automatic gain control, which prevents the output from exceeding a fixed level. The mechanism is physiologically justified by the limited dynamic range of single auditory nerve fibers. For very low sound levels, where the firing rate is barely raised above the spontaneous level, the location of the BMU moves towards silence.

5.3. Gammatone/IHC model

The outcome of the gammatone/IHC model is in many respects similar to that of Lyon's model, because they both attempt to simulate the same processes, using different techniques. In our tests the gammatone/IHC model achieved a more convincing mapping of the test sounds, as shown in Figure 4.

All the tones are arranged in clusters corresponding to their fundamental frequency, but there are no double assignments for the steady state BMUs. Silence is located near the lower left corner, and white noise is, again, on the opposite side. However, it is still not advisable to take the distance between BMUs on the map as a distance measure for the similarity of the sounds, because the overall arrangement is never perfect: the 5000 Hz cluster is located closer to the 100 Hz cluster than to the 500 Hz cluster, the sine octaves 400/800 Hz are mapped next to the 100 Hz tones, etc. The trajectories of the noise signals (01–03) are characterized by an ocsillation between two or three neighbouring units.

As described for Lyon's cochlear model the reduction of the frame rate causes an 'extended reaction time' and smoothes away any short time variations below the 100 Hz frame rate. This applies also to rapid intensity changes produced by the adaptation mechanism. The function of Meddis' IHC model is thus limited to compressing the signal's dynamic range.

5.4. Mel-frequency cepstral coefficients

The distribution of the test sounds on a SOM preprocessed with MFCC is shown in Figure 5. In contrast to the mappings described above, silence is located near the centre of the map, which can be explained by the respective range of values: Lyon's and Meddis' auditory models use only positive values (including zeros for silence), whereas cepstral coefficients can be either positive or negative.

As mentioned above the MFCC trajectories are less smooth than those derived from the low-pass filtered output of the auditory models, and they react to changes in the sound immediately. Trajectories oscillating between two or more units can be found for noise signals as well as for tones. In some cases oscillations extend across larger distances on the map, e.g. for white noise, which overlaps with the 5 kHz square wave. The 100 Hz square wave is mapped to two distinct units because of a phase jump occuring in the middle of the signal, which is reflected by a leap in the trajectory. Unlike the auditory models MFCC preserves this short, but clearly audible event. The quick alterations of noise and tone bursts in sound no. 10 are also shown accurately by the corresponding MFCC trajectory—probably more accurately than by the human ear.

The most important factor in the arrangement of the SOM is, again, the fundamental frequency of the signals. The sound level does not play a significant role, because the power-related coefficient C_0 is discarded, and the other coefficients are largely independent of the sound level. However, when the sound level is very low, the location of the BMU can change remarkably even for small variations.

6. CONCLUSIONS

The functional similarity of the two auditory models in comparison with the MFCC representation is clearly reflected in the resulting SOMs and trajectories, e.g. in the location of the silence-BMU and the 'smoothness' of the trajectories. MFCC is computationally the most efficient representation, but the gammatone filterbank combined with Meddis' inner hair cell model produced the most convincing results on the SOM: the different sounds are clearly separated and still grouped according to their 'similarity'. Similarity is in this case mainly defined by the pitch, or fundamental frequency of the signals, whereas intensity plays only a minor role. Within the 'pitch clusters' the sounds are discriminated on the basis of their timbre.

Further research is required to optimize the output of the auditory models; it would be desirable to preserve part of the temporal information (phase locking) that gets lost during the frame rate reduction. Examples of representations that combine both spectral and temporal information are the *autocorrelogram*, described e.g. by Slaney and Lyon [18], and the *auditory image model*, described by Patterson et al. [8, 9]. However, these models inflate the dimensionality of the data, so that it becomes very expensive for a SOM to process their output.

Future research will include the development of a sequence comparison module to detect similar trajectories by means of string matching algorithms, and the evaluation of the system with more realistic sound examples, such as short pieces of electronic music.

7. ACKNOWLEDGEMENTS

We wish to thank Daniel Teruggi and INA-GRM in Paris for their input to this project and for providing the set of test sounds.

[3] On a larger map it can even take twice as long, because the fine resolution detects even minute deviations from the silence vector.

8. REFERENCES

[1] Bernhard Feiten and Stefan Günzel, "Automatic indexing of a sound database using self-organizing neural nets," *Computer Music Journal*, vol. 18, no. 3, pp. 53–65, 1994.

[2] Petri Toiviainen, "Optimizing self-organizing timbre maps: Two approaches," in *Music, Gestalt, and Computing*, Marc Leman, Ed., pp. 337–350. Springer, Berlin and Heidelberg, 1997.

[3] Petri Toiviainen, Mari Tervaniemi, Jukka Louhivuori, Marieke Saher, Minna Huotilainen, and Risto Näätänen, "Timbre similarity: Convergence of neural, behavioral, and computational approaches," *Music Perception*, vol. 16, no. 2, pp. 223–242, 1998.

[4] Piero Cosi, Giovanni De Poli, and Giampaolo Lauzzana, "Auditory modelling and self-organizing neural networks for timbre classification," *Journal of New Music Research*, vol. 23, no. 1, pp. 71–98, 1994.

[5] Giovanni De Poli and Paolo Prandoni, "Sonological models for timbre characterization," *Journal of New Music Research*, vol. 26, no. 2, pp. 170–197, 1997.

[6] Richard F. Lyon, "A computational model of filtering, detection and compression in the cochlea," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Paris, May 1982.

[7] Malcolm Slaney, "Lyon's cochlear model," Apple Technical Report 13, Apple Computer, 1988, `http://www.slaney.org/malcolm/pubs.html`.

[8] Roy D. Patterson, K. Robinson, John Holdsworth, Denis McKeown, C[elia] Zhang, and Mike Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds., vol. 83 of *Advances in the Biosciences*, pp. 429–445. Pergamon Press, Oxford, 1992, Proceedings of the 9th International Symposium on Hearing held in Carcens, France, June 1991.

[9] Roy D. Patterson and John Holdsworth, "A functional model of neural activity patterns and auditory images," in *Advances in Speech, Hearing and Language Processing*, William A. Ainsworth, Ed., vol. 3. JAI Press, London, 1996, Often cited as Patterson & Holdsworth 1991.

[10] Malcolm Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer Technical Report 35, Apple Computer, 1988, `http://www.slaney.org/malcolm/pubs.html`.

[11] Ray Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *Journal of the Acoustical Society of America*, vol. 79, no. 3, pp. 702–711, Mar. 1986.

[12] Martin Cooke, *Modelling Auditory Processing and Organisation*, Distinguished Dissertations in Computer Science. Cambridge University Press, Cambridge, UK, 1993, PhD thesis, University of Sheffield.

[13] Stephen B. Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuous spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980, Reprinted in [19].

[14] Piero Cosi, Giovanni De Poli, and Paolo Prandoni, "Timbre characterization with mel-cepstrum and neural nets," in *Proceedings of the International Computer Music Conference (ICMC)*, Aarhus, Denmark, 1994, pp. 42–45.

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

[16] Malcolm Slaney, "Auditory Toolbox Version 2," Interval Technical Report 1998-010, Interval Research Corporation, Palo Alto, CA, 1998, `http://www.slaney.org/malcolm/pubs.html`.

[17] Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas, "SOM Toolbox for Matlab 5," Tech. Rep. A57, Helsinki University of Technology, Apr. 2000, `http://www.cis.hut.fi/projects/somtoolbox/`.

[18] Malcolm Slaney and Richard F. Lyon, "On the importance of time: A temporal representation of sound," in *Visual Representations of Speech Signals*, Martin Cooke, Steve Beet, and Malcolm Crawford, Eds., pp. 95–116. John Wiley & Sons, Chichester, UK, 1993.

[19] Alex Waibel and Kai-Fu Lee, Eds., *Readings in Speech Recognition*, Morgan Kaufmann Publishers, San Mateo, CA, 1990.