Computing auditory perception*

HENDRIK PURWINS†, BENJAMIN BLANKERTZ‡ and KLAUS OBERMAYER†

†Neural Information Processing Group FR 2-1, FB 13, Technical University of Berlin, Franklinstr. 28/29, 10587 Berlin, Germany
E-mail: {hendrik,oby}@cs.tu-berlin.de
URL: http://www.cs.tu-berlin.de/~hendrik/
‡GMD-FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
E-mail: blanker@first.gmd.de

In this paper the ingredients of computing auditory perception are reviewed. On the basic level there is neurophysiology, which is abstracted to artificial neural nets (ANNs) and enhanced by statistics to machine learning. There are high-level cognitive models derived from psychoacoustics (especially Gestalt principles). The gap between neuroscience and psychoacoustics has to be filled by numerics, statistics and heuristics. Computerised auditory models have a broad and diverse range of applications: hearing aids and implants, compression in audio codices, automated music analysis, music composition, interactive music installations, and information retrieval from large databases of music samples.

1. INTRODUCTION

What is music? Assume we know in very fine detail how our brain works, while a person performs musical activity. Do we then *understand* music? By this means, we might never understand the subjective quality of an overwhelming musical experience. But we take the view: If we could give a comprehensive analysis of the neural processes related to music activity, we could ground music theory on neurophysiology. We look for general neural principles underlying music activity and for representations of music. Unfortunately, up to now we are far from entirely understanding the brain processes involved. We know a lot about the early auditory processing from outer ear to hair cell. We have some fuzzy knowledge of the localisation of specific abilities in the brain. And we can apply some generally known principles of neural information processing to music. On the whole our comprehension of auditory processing is very fragmentary. How can we fill the gap? In computational neuroscience, neural activity observed in experiments is quantitatively expressed by mathematical formulas. In artificial neural networks, the formalised mechanisms of voltage propagation and neural development principles are abstracted to neuro-mimetic learning algorithms. In this bottom-up approach, algorithms are derived from experiments with single neurons or small neuron populations. These algorithms can perform pattern recognition tasks that resemble human abilities to a low degree.

We can also take a top-down approach. We can observe human performance of auditory activity taken as a whole, by means of psychological experiments. Experiments give rise to hypotheses about underlying cognitive principles that can be manifested by statistical inference. The principles discovered can be turned into computational models. By combining formalised neurons, artificial neural networks and cognitive models, we can implement a model of auditory perception. Can we make use of such a model, e.g. in hearing aids, compression, automated music analysis, or music composition?

We can, to some extent. But often it does not work so well due to the fact that the neurological and psychoacoustic knowledge on which it is based is relatively poor. Often artificial neural networks (ANN) are based on well-known numerical methods with new fashionable names. To outperform classical methods, ANNs require good pre-processing, sensible choice of the network parameters, and many heuristics. Algorithms can be improved by taking into account statistical learning theory (Vapnik 1998). Further improvement is achieved, especially in terms of computational costs, by additional use of numerical techniques, like spectral analysis, filter design, and optimisation.

Hardware implementations of the functionality of the inner ear supply reasonable cochlear implants, whereas implants of later stages of the auditory pathway do not yet aid deaf patients significantly. Implementations of frequency and temporal masking, and just-noticeable differences increase compressibility of audio data. Compressibility is a hot topic in the definition of audio-visual code standards (like MPEG-3,-4,-7), within the convergence of TV/radio, mobile communication, and the Internet. Machine learning algorithms can index sounds by classification. Some ANNs map sounds to a space, in which particular axes represent perceptual qualities. These mappings imply perceptual similarity measurements. Hopefully, with these sound indices and the similarity measurements, large music audio databases can be searched, and information can be retrieved. Filtering music samples according to the auditory periphery, followed by ANNs or statistical analysis with some built-in heuristics, can extract musical parameters such as pitch, harmony, tone centres, metre and rhythm. Taking advantage of competing perceptual rules, paradoxical sounds can be synthesised. These acoustic illusions

* From the ICMC workshop, 'Cognition and Perception Issues in Computer Music' (Ian Whalley and Hendrik Purwins).

correspond to the works of M. C. Escher, and yield interesting compositional effects.

This article has the following structure: In section 2, we take the psychoacoustic top-down view of auditory perception. We start with addressing the problem of identifying an acoustic object within a diversity of auditory impressions by auditory feature binding (the binding problem). We have this capability when separating out a single instrument from an ensemble, during a chamber music performance, for example. Grouping principles such as proximity, good continuation, closure, and common fate are a means of segregating auditory information.

In section 3 we follow the bottom-up approach. We describe the effect of the auditory periphery, the means of information exchange between neurons (spikes and synapses), as a way to calculate amplitude modulations. The tonotopy principle is explained. Two hypotheses regarding the neural implementation of the binding problem are given. Independent component analysis (ICA) and possible applications to the cocktail party problem are briefly introduced.

In section 4, suggestions of perceptual spaces and similarity measures for rhythm, pitch, tone centres and timbre are presented.

Finally, in section 5, it is pointed out how some compositional approaches profit from music cognition and perception: (i) artificial neural nets for automatic composition, (ii) bio-feedback set-ups in interactive installations, and (iii) effective use of auditory illusions.

2. AUDITORY SCENE ANALYSIS

2.1. The binding problem

We are constantly exposed to a chaos of diverse sensory impressions. How can we identify an object in the environment? How can certain sensory impressions form a '*Gestalt*' according to certain criteria and provide us with information about the object in the environment (the 'binding problem')? The 'Gestalt' concept originated from Ehrenfels (1890) and Mach (1886). They initially presented musical examples. Subsequently, visual perception was investigated. From the 1970s on, computer-supported sound synthesis and analysis enforced the application of Gestalt theory to auditory perception, exhaustively reviewed in Bregman (1990).

2.2. Grouping principles

In the following, principles are introduced which aid binding in auditory perception (figure 1, Bregman 1990): the principle of '*proximity*' refers to distances between auditory features with respect to their onsets, pitch and loudness. Features that are grouped together have a small distance between each other, and a long distance to elements of another group. Temporal and pitch proximity are competitive criteria, e.g. the slow sequence of notes A–B–A–B . . . (figure 1, A1), which contains large pitch jumps, is perceived as one stream. The same sequence of notes played very fast (figure 1, A2) produces one perceptual stream consisting of As and another one consisting of Bs.

'*Similarity*' is very similar to proximity, but refers to properties of a sound, which cannot be easily identified with a single physical dimension (Bregman 1990: 198), like timbre.

The principle of '*good continuation*' identifies smoothly varying frequency, loudness or spectra with a changing sound source. Abrupt changes indicate the appearance of a new source. In Bregman and Dannenbring (1973) (figure 1, B), high (H) and low (L) tones alternate. If the notes are connected by glissandi (figure 1, B1), both tones are grouped to a single stream. If high and low notes remain unconnected (figure 1, B2), Hs and Ls each group to a separate stream. 'Good continuation' is the continuous limit of 'proximity'.

The principle of '*closure*' completes fragmentary features, which already have a 'good Gestalt', e.g. ascending and descending glissandi are interrupted by rests (figure 1, C2). Three temporally separated lines are heard one after the other. Then noise is added during the rests (figure 1, C1). This noise is so loud that it would mask the glissando, unless it was interrupted by rests. Amazingly, the interrupted glissandi are perceived as being continuous. They have 'good Gestalt': They are proximate in frequency before and after the rests. So they can easily be completed by a perceived good continuation. This completion can be understood as an auditory compensation for masking.

The principle '*common fate*' groups frequency components together, when similar changes occur synchronously, e.g. synchronous onsets, glides or vibrato. Chowning (1980, figure 1, D) performed the following experiment: First, three pure tones are played; a chord is heard, containing the three pitches. Then the full set of harmonics for three vowels ('oh', 'ah' and 'eh') is added, with the given frequencies as fundamental frequencies, but without frequency fluctuations. This is not heard as a mixture of voices but as a complex sound in which the three pitches are not clear. Finally, the three sets of harmonics are differentiated from one another by their patterns of fluctuation. We then hear three vocal sounds being sung at three different pitches.

Other important topics in auditory perception are *attention* and *learning*. In a cocktail party environment, we can focus on one speaker. Our attention selects this stream. Also, whenever some aspect of a sound changes, while the rest remains relatively unchanging, then that aspect is drawn to the listener's attention ('figure ground phenomenon'). Let us give

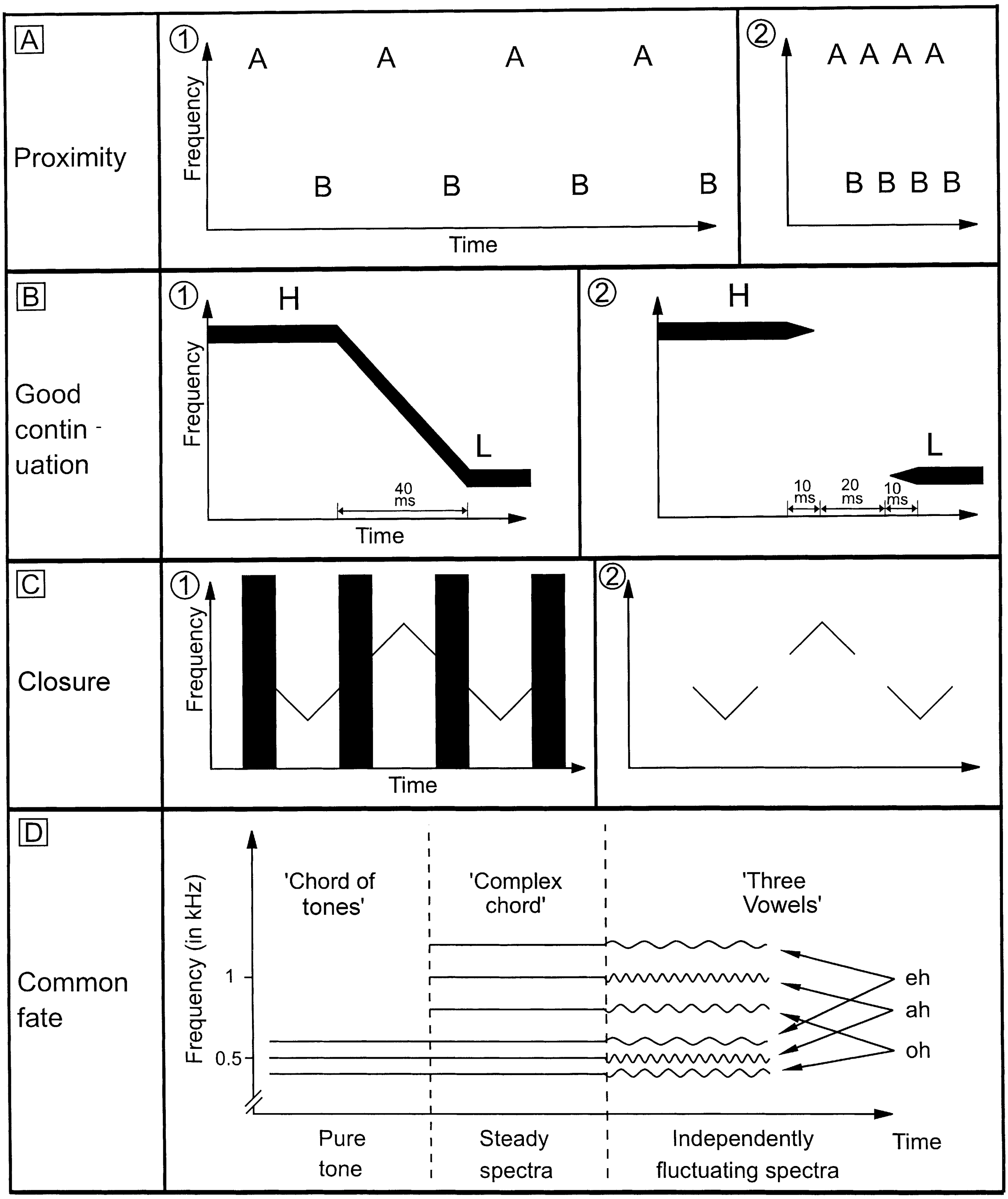


Figure 1. Psychoacoustic experiments demonstrating grouping principles (cf. section 2.2, from Bregman 1990).

an example for learning: the perceived illusory continuity (cf. figure 1, C) of a tune through an interrupting noise is even stronger, when the tune is more familiar (Bregman 1990: 401).

3. COMPUTATIONAL MODELS OF THE AUDITORY PATHWAY

Now the auditory pathway and modelling approaches are described. The common time–log-frequency representation in stave notation originates from the resonating pattern of the basilar membrane. The major properties of the hair cell in the inner ear are temporal coding and adaptation behaviour. An algorithm is presented that implements the auditory principle of tonotopy. Two hypotheses on the neural implementation of the binding problem are approached.

The effect of outer and middle ear can be implemented as a bandpass filter (IIR filter of second order, Oppenheim and Schafer 1989) with a response curve maximal at 2 kHz and decreasing towards lower and higher frequencies.

3.1. Log-frequency coding of the basilar membrane

Incoming sound waves cause a travelling wave on the *basilar membrane*. Hair cells are located on the basilar membrane. Due to varying stiffness of the basilar membrane and the stiffness, size and electrical resonance of the hair cell, different places on the basilar membrane are tuned to different frequencies. Frequencies below 500 Hz are mapped approximately linearly on the basilar membrane. In the range of 50 Hz to 8 kHz the mapping is approximately logarithmic. This is in accordance to the *Weber–Fechner* perceptual law, which is also apparent in the perception of loudness.

This particular mapping has strong implications on pitch perception. The musical understanding of pitch class as octave equivalence is based on uniform spacing of the corresponding resonance frequencies on the basilar membrane. Also, music representation on a music sheet in the time–log-frequency plane reflects that fact. In the range lower than 500 Hz and higher than 8 kHz, the deviation of relative pitch (measured in mel) from strictly logarithmic scaling is due to the resonance properties of the basilar membrane. Apart from correlograms (see below), this time–log-frequency representation is widely used in higher level analysis (see below), e.g. as receptive fields for trajectories in the time–frequency domain (Todd 1999, Weber 2000).

A model of the basilar membrane may be implemented as a filter bank. The spacing and the shape of the filters have to be determined. A first approach is the discrete Fourier transform, which is very quick, but gives equal resolution in the linear (i.e. non-log) frequency domain. For signals which do not contain very low frequencies, the constant Q transform (Brown 1991, Brown and Puckette 1992) is an appropriate method. Equal logarithmic resolution can be achieved also by the continuous wavelet transformation (CWT) (Strang and Nguyen 1997). A more exact modelling of the basilar membrane is supplied by a filter spacing according to the critical band units (CBU) or the equivalent rectangular bandwidths (ERB) (Moore 1989). After mapping frequency on a critical band rate (Bark) scale, masking curves can be approximated by linear functions in the form of a triangle. Gammatone filters realise a fairly good approximation of the filter shape. More complex masking templates are suggested in Fielder, Bosi, Davidson, Davis, Todd and Vernon (1995). Addition of simultaneous masking is an ongoing research topic.

3.2. Information exchange between neurons

Within a neuron, voltage pulses passively propagate along a dendrite until they reach the axon hillock. In the axon hillock, all incoming voltage pulses accumulate until they cross a threshold. As a consequence, a stereotype voltage pulse (*spike*) is generated. The spike propagates across the axon until it reaches a synapse. In the presynaptic axon, neurotransmitters are packed in the vesicles (figure 2(a)). The vesicles release the neurotransmitters into the synaptic cleft, triggered by the incoming spikes. The emitted neurotransmitters enter the receptor channels of the postsynaptic dendrite of the receiving cell.

A sequence of spikes can encode information as follows: (i) by the exact time, when the neuron fires (*time coding hypothesis*), (ii) by the time interval between proceeding spikes, the inter-spike interval, and (iii) the spike rate, the inverse of the mean inter-spike interval. To precisely model (i) and (ii) we could solve a system of partly nonlinear differential equations (Hodgkin and Huxley 1952) describing current flow in the axon. To ease this task we can calculate the 'integrate and fire' model (Maass 1997). Voltage is integrated until threshold is reached. After a refractory period, integration starts again from rest potential. Point (iii) is a rough simplification of spike behaviour and is the basis of the connectionist neuron in artificial neural nets.

According to Meddis and Hewitt (1991), the synapse of the hair cell is formalised as a dynamic system consisting of four elements (figure 2(b)). In this model, the activity transmitted by the hair cell to the auditory nerve is considered proportional to the number of neurotransmitters $c(t)$ in the synaptic cleft. $c(t)$ depends on the number of transmitters $q(t)$ in the hair cell by means of the nonlinear function $k(t)$. $k(t)$ describes the permeability of the presynaptic hair cell membrane and is triggered by the presynaptic membrane potential. The closer $q(t)$ is to the maximum capacity, the less neurotransmitter is produced. A portion of the transmitters (factor r) returns from the synaptic cleft into the presynaptic hair cell. The temporal behaviour of the system is described by a nonlinear first-order differential equation; the change in parameter is calculated from the balance of the incoming and outgoing quantities, e.g.

$$\frac{\partial c}{\partial t} = k(t)q(t) - lc(t) - rc(t).$$

This hair cell model reproduces some experimental results from hair cells of gerbils. In particular, and firstly, frequent spikes occur with the onset of a tone. The spike rate decreases to a constant value, when the tone continues (adaptation behaviour, figure 3(b)). After the offset of the tone, it decreases to about zero. Secondly, below 4–5 kHz, spikes occur in the hair cell (almost) only during the positive phase of the signal (*phase locking*). So in this range, frequency is coded both by the position of the responding hair cell on the basilar membrane and by temporal spike

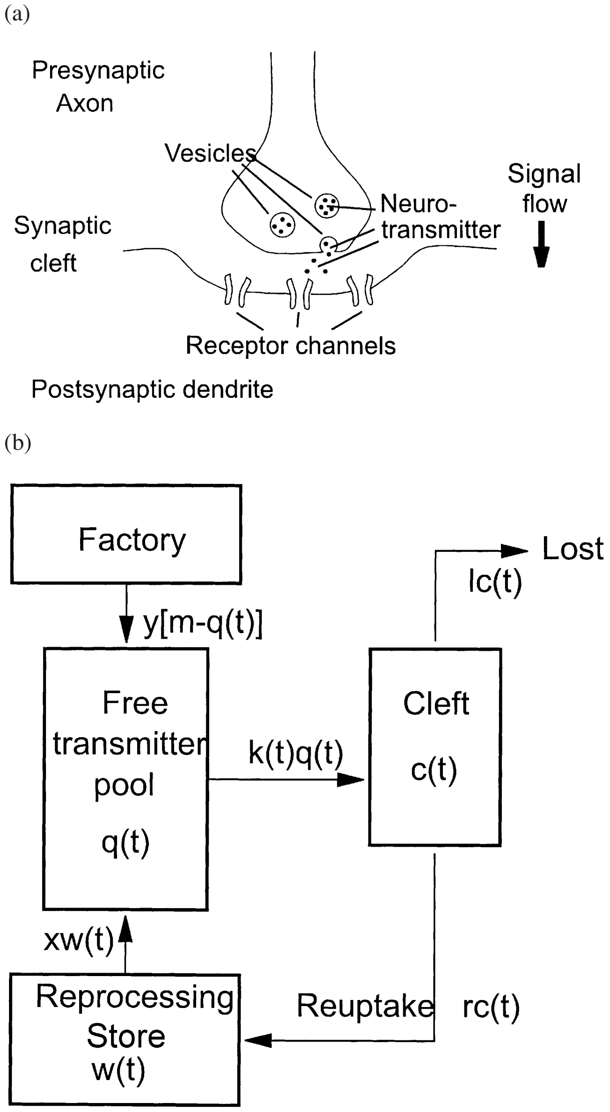


Figure 2. (a) Information is transmitted from the presynaptic to the postsynaptic cell. The action potential in the presynaptic cell forces the vesicles to empty neurotransmitters into the synaptic cleft. The emitted neurotransmitters enter the receptor channels and change the potential of the postsynaptic cell. (b) The hair cell synapse is modelled by a dynamic system consisting of four departments (Meddis and Hewitt 1991). The nonlinear function $k(t)$ includes the presynaptic action potential and controls the permeability of the presynaptic membrane. The evoked potential in the auditory nerve scales with $c(t)$ (cf. section 3.2).

behaviour. For frequencies above 5 kHz, spikes occur about equally often during the positive and the negative phases of the signal. Therefore, above 5 kHz, frequency is only coded by the place information on the basilar membrane.

3.3. Missing fundamental and autocorrelation

Schreiner and Langner (1988) experimentally found neurons that respond to specific amplitude modulation frequencies. They presented some evidence that neurons which respond to increasing amplitude modulation frequencies line up orthogonally to neurons which respond to increasing fundamental frequencies.

Autocorrelation can be used to extract the missing fundamental (Terhardt, Stoll and Seewann 1982) from a complex tone by detecting amplitude modulation. Consider the modified autocorrelation function

$$R_i(n,t) = \sum_k w_t(k) s_i(k) s_i(k+n),$$

where $w_t(k)$ is a rectangular window possibly multiplied by an exponential function. Let $R_n(t)$ be the summation of $R_i(n,t)$ across all channels i. This means that the basic periodicities of the signal and their multiples are added up in each channel i. A maximum in $R_n(t)$ corresponds to the least common multiple of the basic periods of the single frequency components. In terms of frequencies, this represents the greatest common divisor. This corresponds fairly well to the virtual pitch of the signal (figure 3(c)).

Scheirer (1999) considers an autocorrelation–time representation ('correlogram') a more perceptually plausible representation than a frequency–time representation. He gives some examples of segregation of different instruments and voices extracted from a mixture. However, the biological plausibility of autocorrelation is still subject to discussion.

3.4. Later auditory processing

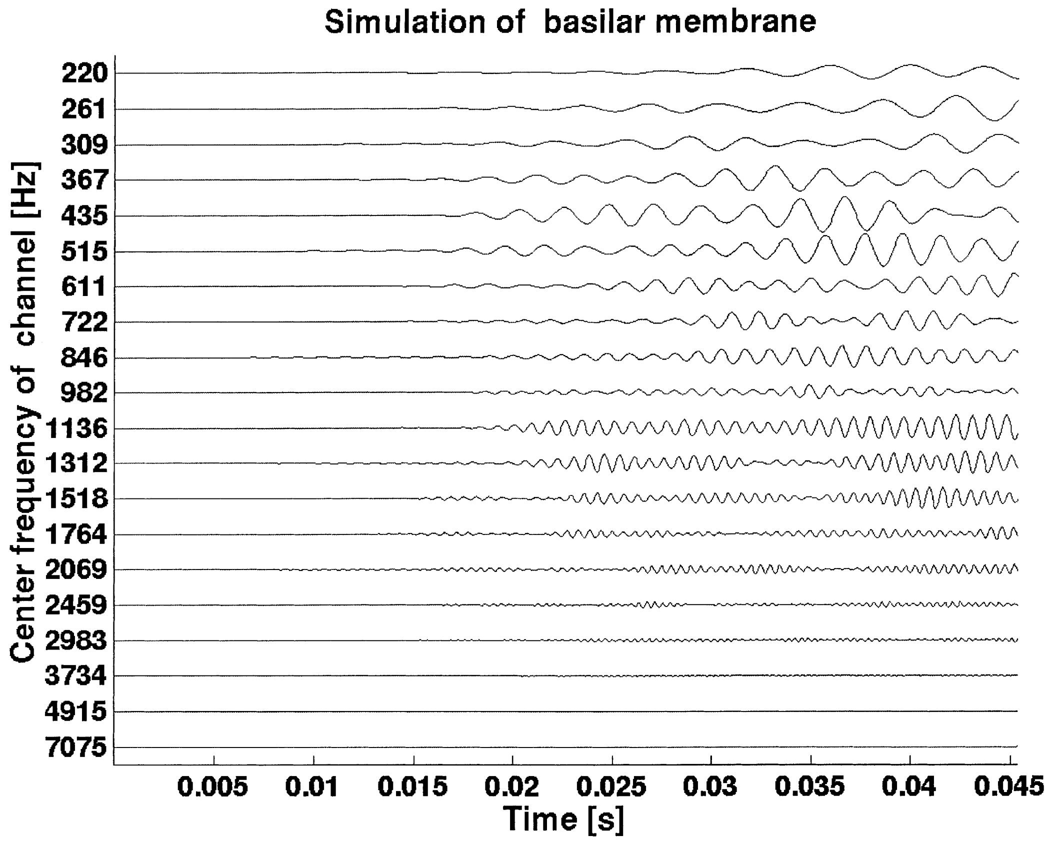
Following the hair cells, the signal chain is continued by the auditory nerve, the auditory nucleus, the superior olivary nucleus, the inferior colliculus, the medial geniculate nucleus, the primary auditory cortex, and higher areas. Knowledge about music processing in the auditory pathway subsequent to the hair cell, especially in the cortex, is poor. In addition to invasive methods in animals, methods for investigation in humans comprise functional magnetic resonance imaging (fMRI), electroencephalogram (EEG, e.g. using event-related potentials, ERP), or implanted electrodes in epileptic patients. In the auditory nucleus, cells are found that response to onset and offset tones. Cells in the superior olivary nucleus help spatial localisation of sound sources by inter-aural phase and intensity differences. In cortex, some vague differences of brain activity are observed, according to different music styles or musical training of the test subjects (Petsche 1994, Marin and Perry 1999).

There are also feedback connections that tune the auditory periphery. It is observed that the cochlea and middle ear can be activated from higher areas to *produce* sounds (*otoacoustic emissions*).

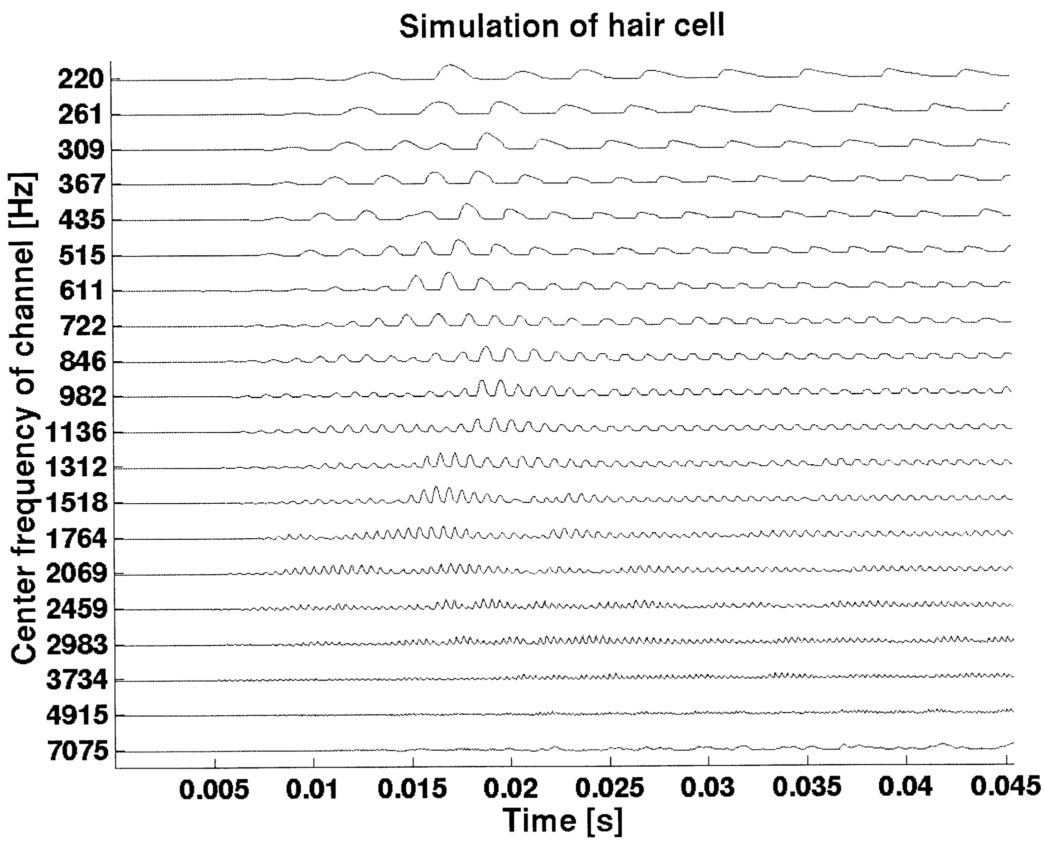
3.5. Hebbian learning

If presynaptic and postsynaptic electric activity occur synchronously, the postsynaptic receptor channels

(a)


Simulation of basilar membrane
Center frequency of channel [Hz]
220
261
309
367
435
515
611
722
846
982
1136
1312
1518
1764
2069
2459
2983
3734
4915
7075
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
Time [s]

(b)


Simulation of hair cell
Center frequency of channel [Hz]
220
261
309
367
435
515
611
722
846
982
1136
1312
1518
1764
2069
2459
2983
3734
4915
7075
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
Time [s]

(c)

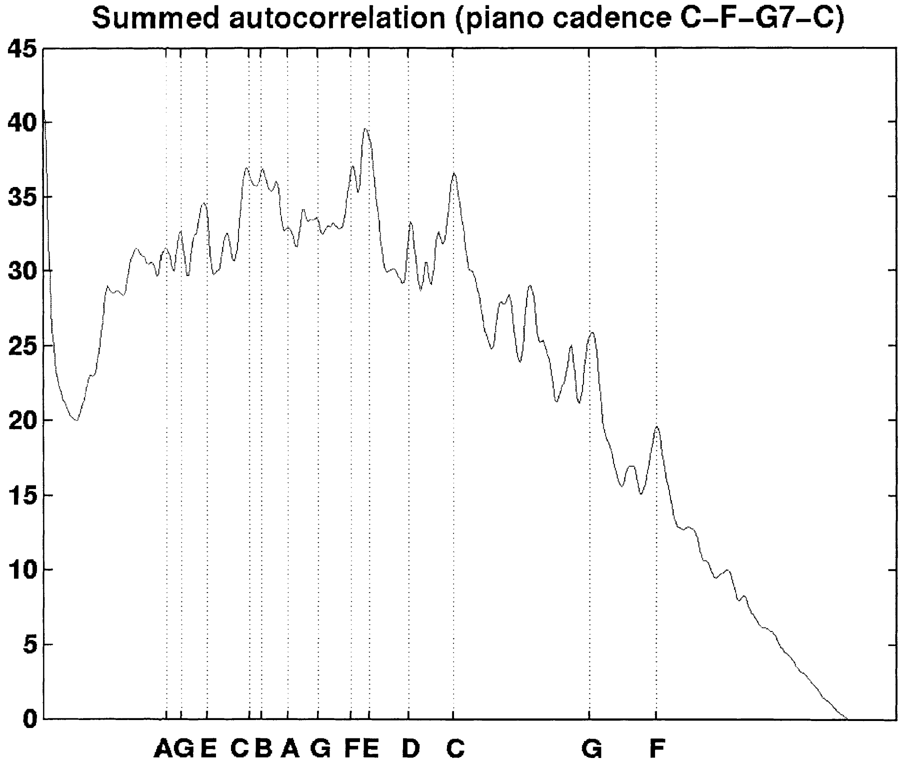


Figure 3. A cadential chord progression (C–F–G7–C) played on the piano is processed by (a) an auditory model consisting of a basilar membrane filter bank, (b) a hair cell model (Meddis and Hewitt 1991), and (c) autocorrelation and temporal integration. Tones corresponding to peaks in the correlogram are indicated (cf. section 3.1–3).

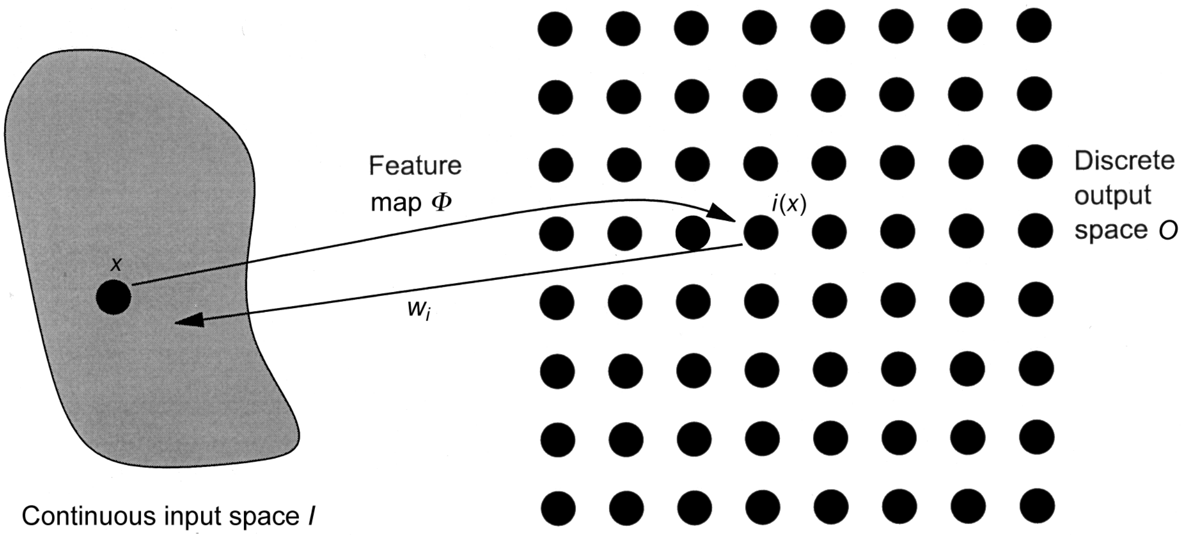


Figure 4. The self-organising feature map (Kohonen 1982) realises the tonotopy principle in the auditory pathway. Φ is a nonlinear mapping between the continuous somato-sensory input space I and the discrete 'cortex' output space O. Given an input vector x, first a best-matching neuron $i(x)$ in O is identified. The synaptic weight vector w_i of neuron $i(x)$ may be viewed as the coordinates of the image of neuron i projected in the input space (from Haykin 1999).

become more permeable, so that a presynaptic activity evokes stronger activity on the postsynaptic dendrite. This principle is called *Hebbian Learning*.

According to the principle of *tonotopy*, proximate hair cells on the basilar membrane project to proximate neurons in the central nervous system. In computer science, the tonotopic principle is realised by an algorithm, the 'self-organising feature map' (SOM, Kohonen 1982, figure 4). The tonotopy property of the SOM is also optimal in the sense of the information theoretic principle of 'maximal information preservation' (Linsker 1989). A more noise-robust version of the SOM is given by Graepel, Burger and Obermayer (1997).

3.6. Feature binding by hierarchical organisation

A hypothetical solution to the binding problem works via integration by anatomic convergence. This model assumes that at an early stage, basic object features such as frequency components are detected. Through progressive convergence of the connections, cells emerge with more specific response properties on a higher processing level. For example, they respond to tones, chords, harmonies and keys. This corresponds to hierarchical artificial intelligence approaches (cf. context-free grammars). Even though hierarchical structures in the brain are joined by lateral connections, in practice a strictly hierarchical concept is successful, e.g. within a framework of a knowledge database and a Bayesian network (Kashino, Nakadai, Kinoshita and Tanaka 1998).

3.7. Feature binding by neural synchronisation

Another way of trying to explain feature binding is through neural synchronisation. The temporal binding model assumes that assemblies of synchronously firing neurons represent objects in the cortex. For example, such an assembly would represent a particular speaker. These assemblies comprise neurons, which detect specific frequencies or amplitude modulation frequencies. The relationship between the partials can then be encoded by the temporal correlation among these neurons. The model assumes that neurons, which are part of the same assembly, fire in synchrony, whereas no consistent temporal relation is found between cells belonging to representations of different speakers. Evidence for feature binding by neural synchronisation in the visual cortex is given by Engel, Roelfsema, Fries, Brecht and Singer 1997).

Terman and Wang (1995), Wang (1996) and Brown and Cooke (1998) supply an implementation based on time–log-frequency representation of music. Their model consists of a set of oscillators in the time–frequency domain or the correlogram. Oscillators which are close to each other are coupled strongly (Hebbian rule and principle of proximity). An additional global inhibitor stops oscillators belonging to different streams being active at the same time. This approach can be used for vowel segregation and also for segregation of different voices according to the proximity principle (Cooke and Brown 1999, figure 1, A). A more promising approach is based on the 'integrate and fire' model (cf. section 3.2, Maass 1997). An ensemble of such models displays synchronous spike patterns.

3.8. Implementations of acoustical source separation

A mixture of several sound sources (speech, music, other sounds) is recorded by one or several microphones. The aim is the decomposition into the original sources. Since an important application is the development of hearing aids, the goal is demixing with at most two microphones.

There are some successful applications of source separation for artificial mixtures (sources are recorded separately and are then digitally mixed by weighted addition). On the other hand, mixtures in real environments are more difficult to demix. The different approaches can be roughly divided into two categories: (i) mimicking the auditory system, (ii) employment of techniques of digital signal processing without reference to biology. Okuno, Ikedo and Nakatani (1999) aims to combine (i) and (ii) synergetically. A possible treatment similar to (i) is as follows. An auditory model is used for pre-processing and from the output of the auditory model (cochleagrams and correlograms), harmonic substructures are extracted. By the use of Gestalt principles, spectral units are built from this. From these separated units, sound can be resynthesised (Nakatani, Okuno and Kawabata 1995). In another approach the spectral units are determined as a sequence of correlograms, and the auditory transformation is inversely calculated (Slaney 1994, 1998).

3.9. Independent component analysis

The larger portion of methods according to (ii) above deals with the realisation of the *Independent Component Analysis* (ICA, Comon 1994, Cardoso 1998, Müller, Philips and Ziehe 1999). Sanger (1989) indicates a reference to biological systems, but his approach is largely a purely statistical model. Ideally, the problem is modelled as follows. The sources are transformed into m sensor signals $x_1(t), \ldots, x_m(t)$ by temporally constant linear mixtures. The sensor signals are the signals recorded by the different microphones (t is the time index, which is omitted in the sequel). With $s = (s_1, \ldots, s_n)^T$ and $x = (x_1, \ldots, x_m)^T$, we can put this in matrix notation as $x = As$, where A denotes the (unknown) mixing matrix. We aim at identifying the demixing matrix W, so that for $y = Wx$ the components of y correspond to the source signals s. (In principle, order and scaling of the rows of s cannot be determined.) The ICA approach to solve this problem is based on the assumption that the source signals s_i are distributed statistically independent (and at most one is Gaussian). The determination of the demixing matrix is possible if there are at least as many microphones as sources, and the mixing matrix A is invertible. In practical applications, the assumptions of statistical independence and the invertibility of A are not critical. Nevertheless, it is problematic that the model does not account for real reverberation. So decomposition of mixtures in a real acoustic environment works only under very special conditions. First approaches are by Lee, Girolami, Bell and Sejnowski (1998), Casey and Westner (2000), Parra and Spence (2000) and Murata, Ikeda and Ziehe (2000). In addition, it is still a tough problem to separate one speaker from a cocktail party environment with several sound sources using only two microphones, as all applicable algorithms up to now require as many microphones as sound sources.

4. GEOMETRIC MODELS

We can map sounds on *perceptual spaces*. The axes of the space correspond to perceptual proximities, in particular musical parameters. There are geometrical perceptual models of pitch, keys, timbre and emotions. Applications are found in composition, information retrieval and music theory.

Camurri, Coletta, Ricchetti and Volpe (2000) use a geometrical arrangement of emotional states combined with a mapping of emotions to rhythmic and melodic features. In their installation, autonomous robots react to visitors by humming sad or happy music, and by moving in a particular manner. Whalley (2000a, b) models the psychological interplay of emotions in a character as a physical dynamic system, whose parameters are mapped to musical entities. The dynamic system determines the underlying structure of the piece.

Geometric models are also used to visualise proximities of single musical parameters: Shepard (1982) supplies a geometric model of pitch: the perceived pitch distance is calculated as the combined Euclidean distance of the notes on the pitch height axis, the chroma circle, and the circle of fifths. For building geometric models of perception, visualisation and clustering algorithms are often used, such as multi-dimensional scaling (MDS, Shepard 1962), principal component analysis (ICA, Comon 1994, cf. section 3.9), or SOM (Kohonen 1982, cf. section 3.5 and figure 4). For example, a plausible model of inter-key relations is provided by the MDS analysis of psychoacoustic experiments (Krumhansl and Kessler 1982): all major and minor keys are arranged on a torus, preserving dominant, relative and parallel major/minor relations.

This structure also emerges from a learning procedure with real music pieces as training data (Leman 1995, Purwins, Blankertz and Obermayer 2000). The constant Q transform (Brown and Puckette 1992, Izmirli and Bilgen 1996) can be applied, to calculate the *constant Q profile* (Purwins *et al.* 2000, figure 5). The constant Q analysis is used as a simple cognitive model. The only explicit knowledge that is incorporated in the model is octave equivalence and well-tempered tuning. This represents minimal information about tonal music. The relevant inter-key relations (dominant, relative, and parallel relations) emerge merely on the basis of trained music examples (Chopin's Préludes, op. 28) in a recording in audio data format (figure 6).

Since the perception of timbre is complex, it is challenging to project timbres into a two- or three-dimensional timbre space, in which the dimensions are related to physical quantities (Grey 1977, McAdams, Winsberg, Donnadieu, DeSoete and Kimphoff 1995, Lakatos 2000). For example, in the three-dimensional timbre space of McAdams *et al.* (1995), the first dimension is identified with the log attack time, the second with the harmonic spectral centroid, and the third with a combination of harmonic spectral spread and harmonic spectral variation.

By using appropriate interpolation, composers can build synthesis systems which are triggered by these models: moving in timbre space can yield sound morphing; walking on the surface of the tone centre torus results in a smooth modulation-like key change. The implied similarity measures of the geometrical models aid information retrieval in huge sound databases, e.g. searching for a music piece, given a melody.

5. COGNITION AND PERCEPTION IN COMPOSITION

5.1. Automated composition

The use of neuro-mimetic artificial neural nets for music composition is only partially successful. Melody generation with the 'backpropagation through time' algorithm with perceptually relevant pre-processing (Mozer 1994), as well as using stochastic Boltzmann machines for choral harmonisation (Bellgard and Tsang 1994) does not yield musically pleasant results. They do not reach the quality of compositions generated by the elaborated rule-based system in Cope (1996). Bach choral harmonisation with a simple feed-forward net (Feulner 1993), and harmonisation in real time (Gang and Berger 1997) with a sequential neural network are more musically interesting.

5.2. Biofeedback music installations

It is possible to find some biophysical correlates of emotional content by measurements of cardiac (ECG), vascular, electrodermal (GSR), respiratory, and brain functions (EEG) (Krumhansl 1997). Biophysical measurement can be used in a biofeedback set-up in a music performance. Constantly biophysical functions are recorded from the performer. These data are used to control music synthesis. In EEG, detection of alpha waves (Knapp and Lustad 1990), or state changes interpreted as attention shifts (Rosenboom 1990) are clues to generate musical structure in real time. During such a feedback set-up, control of skin conductance and heartbeat can be learned by the performer.

5.3. Psychoacoustic effects

Stuckenschmidt (1969) reports that Ernst Krenek electronically generated sounds using a method similar to the

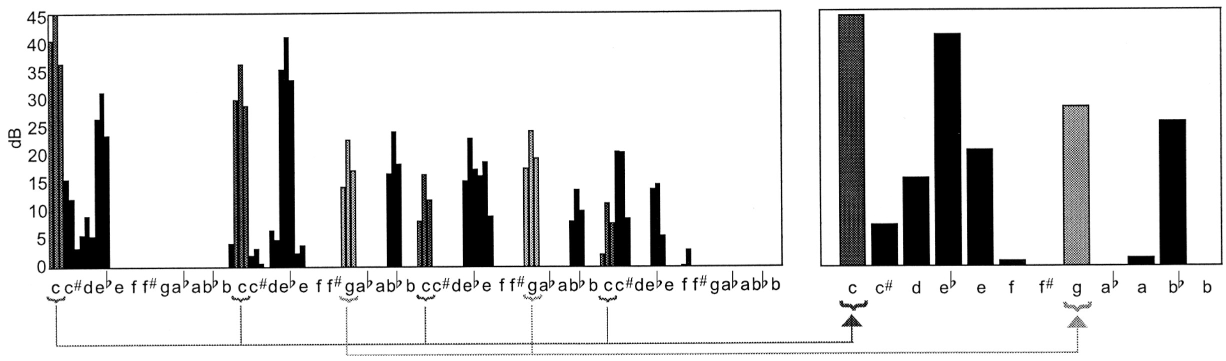


Figure 5. The constant Q transform (Brown and Puckette 1992) and the cq-profile (Purwins, Blankertz and Obermayer 2000) of a minor third (C-E♭) played on a piano. A twelve-dimensional constant Q profile is closely related to pitch classes and to probe tone ratings (Krumhansl and Kessler 1982). Cq-profiles can be efficiently calculated; they are stable with respect to sound quality of the music sample and they are transposable (cf. sections 3.1 & 3).

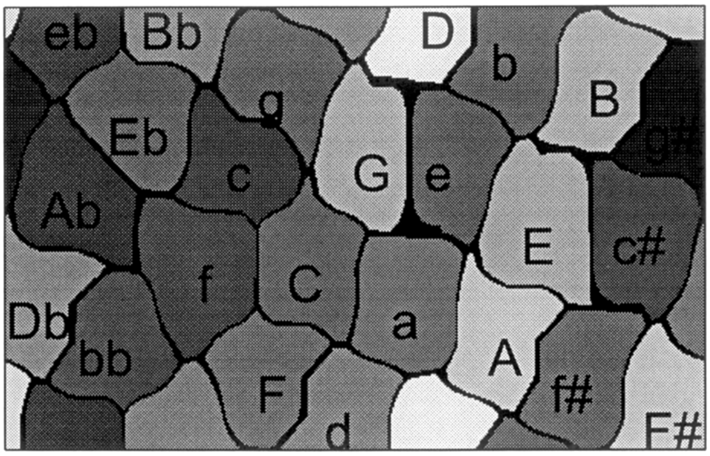


Figure 6. Emergence of inter-key relations. Inter-key relations are derived from a set-up including constant Q profile calculation and a toroidal SOM (cf. figure 4) trained by Chopin's Préludes, op. 28 recorded by A. Cortot in 1932/33, in audio data format. The image shows a torus surface. Upper and lower, left and right sides are to be glued together. Keys with dominant, or major/minor parallel and relative relations are proximate (cf. section 4).

endlessly ascending or descending Shepard scales in the oratorio *Spiritus Intelligentiae Sanctus* (1955–6) to create a sense of acoustic infinity. Risset (1985) used many musical illusions in his pieces. He extended the Shepard scales to gliding notes and invented endlessly accelerating rhythmic sequences. Conflicts between groupings based on timbre similarity on the one hand and on pitch proximity on the other can create musical form. In the instrumentation of Bach's Ricercar from *The Musical Offering*, Webern outlines musical gestures and phrases by changing instrumentation after each phrase. The particular phrases are focused, yet the melody is still recognisable (Bregman 1990: 470). There are still a couple of perceptual effects in auditory scene analysis to be explored for compositional use, even though a good composer might intuitively know about them already.

6. CONCLUSION

Do findings from neuroscience, psychoacoustics, music theory, and computational models match? Grouping principles aid the understanding of voice leading and harmonic fusion. Stave notation, the widely used representation of music, stems from the frequency coding on the basilar membrane. Classical ANNs are based on mere spike rate, whereas current research seems to support the time coding hypothesis. The specific timing of spikes transmits information. Time coding would enable a neural implementation of binding by synchronised spiking. The hair cell synapse model (Meddis and Hewitt 1991) reveals adaptation behaviour, which correlates to the way attention is directed in an auditory scene. Pace and temporal coding of frequency correspond to relative pitch perception (mel) in the frequency range above 5 kHz. We suggest that a computational model can easily mimic any required behaviour, by introducing numerous parameters. That would contradict Occam's razor favouring the simplest model. The connection between perception of virtual pitch and neurons sensitive to auditory amplitude modulations is not yet entirely understood. SOM and ICA represent a high abstraction level from neurobiology. In this paper we could have additionally described the effective support vector machine (SVM, Smola and Schölkopf 1998) for regression and classification, Bayesian networks, or hidden Markov models (HMM, Rabiner 1989), which yield good results when applied to pre-processed language. However, these models have even less biological plausibility.

Progress in neuroscience is developing rapidly. It has great influence on our field. All sorts of applications entirely based on a thorough knowledge of the auditory system might be used to good advantage. However, we have to recognise that music composition is based on perceptual rules, rather than on pure mathematics and rough music aesthetics.

ACKNOWLEDGEMENTS

The first author was supported by the German National Merit Foundation and the Axel Springer Stiftung. He is indebted to Al Bregman and Pierre Ahad for their hospitality at McGill University. Special thanks go to Ross Kirk and Gregor Wenning.

WEB RESOURCES

Cooke, M., and Brown, G. 1998. Matlab Auditory Demonstrations (MAD), Version 2.0: http://www.dcs.shef.ac.uk/~martin/MAD/docs/mad.htm (good demonstrations of oscillators, runs on UNIX).

O'Mard, L. P. 1997. Development System for Auditory Modelling (DSAM), Version 2.0: http://www.essex.ac.uk/psychology/hearinglab/dsam/home.html (large, requires some effort to learn).

Slaney, M. 1998. Auditory Toolbox for Matlab, Version 2: http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/ (portable, easy to handle, useful for frequency estimation).

Auditory mailing list archive including actual research topics: http://sound.media.mit.edu/dpwe-bin/mhindex.cgi/AUDITORY/postings/2000

Information resource for auditory neurophysiology: http://neuro.bio.tu-darmstadt.de/langner/langner.html

RECOMMENDED TEXTBOOKS AND COLLECTIONS

Bregman, A. S. 1990. *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

Hawkins, H. L., McMullen, T. A., Popper, A. N., and Fay, R. R. (eds.) 1996. *Auditory Computation*, Springer Handbook of Auditory Research 6. New York: Springer.

Haykin, S. 1999. *Neural Networks*, 2nd edn. Upper Saddle River, NJ: Prentice-Hall.

Kandel, E. R., Schwartz, J. H., and Jessell, T. M. (eds.) 1991. *Principles of Neural Science*, 3rd edn. Norwalk, CT: Appleton & Lange.

Moore, B. C. J. 1989. *An Introduction to the Psychology of Hearing*, 3rd edn. London: Academic Press.

Oppenheim, A. V., and Schafer, R. W. 1989. *Discrete-time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Rosenthal, D. F., and Okuno, H. G. (eds.) 1998. *Computational Auditory Scene Analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Strang, G., and Nguyen, T. 1997. *Wavelets and Filter Banks*. Wellesley, MA: Wellesley-Cambridge Press.

OTHER REFERENCES

Bellgard, M. I., and Tsang, C. P. 1994. Harmonizing music the Boltzmann way. *Connection Science* **6**(2&3): 281–97.

Bregman, A. S., and Dannenbring, G. 1973. The effect of continuity on auditory stream segregation. *Perception Psychophysics* **13**: 308–12.

Brown, G. J., and Cooke, M. 1998. Temporal synchronization in a neural oscillator model of primitive auditory stream segregation. In D. F. Rosenthal and H. G. Okuno (eds.) *Computational Auditory Scene Analysis*, pp. 87–103. Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, J. 1991. Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America* **89**(1): 425–34.

Brown, J., and Puckette, M. S. 1992. An efficient algorithm for the calculation of a constant Q transform. *Journal of the Acoustical Society of America* **92**(5): 2,698–701.

Camurri, A., Coletta, P., Ricchetti, M., and Volpe, G. 2000. Synthesis of expressive movement. *Proc. ICMC*, pp. 270–3. International Computer Music Association.

Cardoso, J.-F. 1998. Blind signal separation: statistical principles. In *Proc. of the IEEE, Special Issue on Blind Identification and Estimation*.

Casey, M. A., and Westner, A. 2000. Separation of mixed audio sources by independent subspace analysis. *Proc. ICMC*, pp. 154–61. International Computer Music Association.

Chowning, J. M. 1980. Computer synthesis of the singing voice. *Sound Generation in Winds, Strings, Computers*, p. 29. Stockholm: Royal Swedish Academy of Music.

Comon, P. 1994. Independent component analysis, a new concept? *Signal Processing* **36**(3): 287–314.

Cooke, M. P., and Brown, G. J. 1999. Interactive explorations in speech and hearing. *Journal of the Acoustical Society of Japan* **20**(2): 89–97.

Cope, D. 1996. *Experiments in Musical Intelligence*. Madison, WI: A-R Editions.

Ehrenfels, C. von. 1890. Über Gestaltqualitäten. *Vierteljahresschrift Wiss. Philos.* **14**: 249–92.

Engel, A., Roelfsema, P. R., Fries, P., Brecht, M., and Singer, W. 1997. Role of the temporal domain for response selection and perceptual binding. *Cerebral Cortex* **7**: 571–82.

Feulner, J. 1993. Neural networks that learn and reproduce various styles of harmonization. *Proc. ICMC*, pp. 236–8. International Computer Music Association.

Fielder, L. D., Bosi, M., Davidson, G., Davis, M., Todd, C., and Vernon, S. 1995. AC-2 and AC-3: low-complexity transform-based audio coding. In *Collected Papers on Digital Audio Bit-Rate Reduction*.

Gang, D., and Berger, J. 1997. A neural network model of metric perception and cognition in the audition of functional tonal music. In *Proc. ICMC*. International Computer Music Association.

Graepel, T., Burger, M., and Obermayer, K. 1997. Phase transitions in stochastic self-organizing maps. *Physical Review E*. **56**(4): 3,876–90.

Grey, J. M. 1977. Multidimensional perceptual scaling of musical timbre. *Journal of the Acoustical Society of America* **61**: 1,270–7.

Hodgkin, A. L., and Huxley, A. F. 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology* **117**: 500–44.

Izmirli, Ö., and Bilgen, S. 1996. A model for tonal context time course calculation from acoustical input. *Journal of New Music Research* **25**(3): 276–88.

Kashino, K., Nakadai, K., Kinoshita, T., and Tanaka, H. 1998. Application of the Bayesian probability network to music scene analysis. In D. F. Rosenthal and H. G. Okuno (eds.) *Computational Auditory Scene Analysis*, pp. 115–37. Mahwah, NJ: Lawrence Erlbaum Associates.

Knapp, B. R., and Lustad, H. 1990. A bioelectric controller for computer music applications. *Computer Music Journal* **14**(1): 42–7.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43**: 59–69.

Krumhansl, C. L. 1997. Psychophysiology of musical emotions. *Proc. ICMC*, pp. 3–6. International Computer Music Association.

Krumhansl, C. L., and Kessler, E. J. 1982. Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review* **89**: 334–68.

Lakatos, S. 2000. A common perceptual space for harmonic and percussive timbres. *Perception and Psychophysics* (in press).

Lee, T.-W., Girolami, M., Bell, A. J., and Sejnowski, T. J. 1998. A unifying information-theoretic framework for independent component analysis. *International Journal on Mathematical and Computer Modeling*.

Leman, M. 1995. *Music and Schema Theory*, Springer Series in Information Sciences 31. Berlin: Springer.

Linsker, R. 1989. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation* **1**: 402–11.

Maass, W. 1997. Networks of spiking neurons: the third generation of neural network models. *Neural Networks* **10**: 1,659–71.

Mach, E. 1886. *Beiträge zur Analyse der Empfindungen*. Jena.

Marin, O. S. M., and Perry, D. W. 1999. Neurological aspects of music perception and performance. In D. Deutsch (ed.) *The Psychology of Music*, 2nd edn, pp. 653–724. Academic Press Series in Cognition and Perception. San Diego: Academic Press.

McAdams, S., Winsberg, S., Donnadieu, S., DeSoete, G., and Kimphoff, J. 1995. Perceptual scaling of synthesized musical timbres: common dimensions, specificities and latent subject classes. *Psychological Research* **58**: 177–92.

Meddis, R., and Hewitt, M. J. 1991. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America* **89**(6): 2,866–82.

Mozer, M. C. 1994. Neural network music composition by prediction: exploring the benefits of psychoacoustic constraints and multi-scale processing. *Connection Science* **6**(2&3): 247–80.

Müller, K.-R., Philips, P., and Ziehe, A. 1999. JADE-TD: combining higher-order statistics and temporal information for blind source separation (with noise). *Proc. of the First Int. Workshop on Independent Component Analysis and Signal Separation: ICA'99*, pp. 87–92. Assios, France.

Murata, N., Ikeda, S., and Ziehe, A. 2000. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputation* (in press).

Nakatani, T., Okuno, H. G., and Kawabata, T. 1995. Residue-driven architecture for computational auditory scene analysis. *Proc. of the 14th Int. Joint Conf. on Artificial Intelligence (IJCAI-95)*, pp. 165–72.

Okuno, H. G., Ikeda, S., and Nakatani, T. 1999. Combining independent component analysis and sound stream segregation. *Proc. of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99)*, pp. 92–8.

Parra, L., and Spence, C. 2000. Convolutive blind separation of non-stationary sources. *IEEE Transactions on Speech and Audio Processing*, pp. 320–7.

Petsche, H. 1994. Zerebrale Verarbeitung. In H. Bruhn, R. Oerter and H. Rösing (eds.) *Musikpsychologie*, pp. 630–8. Reinbek: Rowohlts Enzyklopädie.

Purwins, H., Blankertz, B., and Obermayer, K. 2000. A new method for tracking modulations in tonal music in audio data format. *Int. Joint Conf. on Neural Networks 2000* **6**: 270–5. IEEE Computer Society.

Rabiner, L. R. 1989. A tutorial on hidden Markov models. *Proc. of the IEEE* **73**: 1,349–87.

Risset, J.-C. 1985. Computer music experiments 1964– . . . *Computer Music Journal* **9**(1): 11–18.

Rosenboom, D. 1990. The performing brain. *Computer Music Journal* **14**(1): 48–66.

Sanger, T. D. 1989. Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks* **2**: 459–73.

Scheirer, E. D. 1999. Towards music understanding without separation: segmenting music with correlogram comodulation. *MIT Media Laboratory Perceptual Computing Section Technical Report*, No. 492.

Schreiner, C. E., and Langner, G. 1988. Coding of temporal patterns in the central auditory nervous system. In G. M. Edelman, W. Gall and W. Cowan (eds.) *Auditory Function: Neurobiological Bases of Hearing*, pp. 337–62. New York: John Wiley and Sons.

Shepard, R. N. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika* **2**(27): 125–40.

Shepard, R. N. 1982. Geometrical approximations to the structure of musical pitch. *Psychological Review* **89**: 305–33.

Slaney, M. 1994. Auditory model inversion for sound separation. *ICASSP*. Adelaide, Australia.

Smola, A. J., and Schölkopf, B. 1998. A tutorial on support vector regression. NeuroCOLT2. *Technical Report Series*, NC2-TR-1998-030.

Stuckenschmidt, H. H. 1969. *Twentieth Century Music*. New York: McGraw-Hill.

Terhardt, E., Stoll, G., and Seewann, M. 1982. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *Journal of the Acoustical Society of America* **71**(3): 679–88.

Terman, D., and Wang, D. L. 1995. Global competition and local cooperation in a network of neural oscillators. *Physica D* **81**: 148–76.

Todd, N. 1999. Implications of a sensory-motor theory for the representation and segregation of speech. *Journal of the Acoustical Society of America* **105**(2): 1,307.

Vapnik, V. 1998. *Statistical Learning Theory*. New York: Jon Wiley and Sons.

Wang, D. L. 1996. Primitive auditory segregation based on oscillatory correlation. *Cognitive Science* **20**: 409–56.

Weber, C. 2000. *Maximum a Posteriori Models for Cortical Modeling: Feature Detectors, Topography and Modularity*. Ph.D. thesis, Berlin.

Whalley, I. 2000a. Emotion, theme and structure: enhancing computer music through system dynamics modelling. *Proc. ICMC*, pp. 213–6. International Computer Music Association.

Whalley, I. 2000b. Applications of system dynamics modelling to computer music. *Organised Sound* **5**(3): 149–57.