Computational Models of Music Perception and Cognition I: The Perceptual and Cognitive Processing Chain (*Preprint*)

Hendrik Purwins[1], Perfecto Herrera[1], Maarten Grachten[1,2], Amaury Hazan[1], Ricard Marxer[1], and Xavier Serra[1]

[1]Music Technology Group
Universitat Pompeu Fabra, Barcelona

[2] Department of Computational Perception
Johannes Kepler Universität, Linz [1]

Address of Correspoding Author:
Universitat Pompeu Fabra
Institut Universitari de Audiovisual
Music Technology Group
Ocata 1
08003 Barcelona, Spain
Tel: 0034-93 54 21 365
Email: hpurwins@iua.upf.es

PACS: 43.75.Cd Music perception and cognition
43.75.St Musical performance, training, and analysis
43.75.Xz Automatic music recognition, classification, and information retrieval
43.75.Zz Analysis, synthesis, and processing of musical sounds

Keywords: music cognition, auditory system, music perception, musical expectancy

[1] Please cite as : Purwins et al., Computational Models of Music Perception and Cognition I: The Perceptual and Cognitive Processing Chain, Physics of Life Reviews. 5(3), 151-168.

Contents

Abstract

We present a review on perception and cognition models designed for or applicable to music. An emphasis is put on computational implementations. We include findings from different disciplines: neuroscience, psychology, cognitive science, artificial intelligence, and musicology. The article summarizes the methodology that these disciplines use to approach the phenomena of music understanding, the localization of musical processes in the brain, and the flow of cognitive operations involved in turning physical signals into musical symbols, going from the transducers to the memory systems of the brain. We discuss formal models developed to emulate, explain and predict phenomena involved in early auditory processing, pitch processing, grouping, source separation, and music structure computation. We cover generic computational architectures of attention, memory, and expectation that can be instantiated and tuned to deal with specific musical phenomena. Criteria for the evaluation of such models are presented and discussed. Thereby, we lay out the general framework that provides the basis for the discussion of domain-specific music models in Part II.

1 Introduction

The faculty of music seems to have a special and, at present, mysterious status for humans. As is the case with language, music understanding and production is nearly exclusively found in humans. As with language, specific areas of the brain seem to be devoted to the processing of music information. If we could grasp universal principles of musical intelligence, we would get an idea of how our music understanding gets refined and adapted to a particular musical style as a result of a developmental process triggered by stimuli of that musical culture. Traditional music theory [Rameau, 1722; Riemann, 1877; Schenker, 1935] tries to find universal laws of music. But when applied to various styles of music the limitations of these theories become apparent. Each theoretic approach is more or less adapted to the musical style it describes best (Purwins [2005] p. 21). Traditional music theory can be used as a guide, but not as a *normative* reference that dictates the kind of processes and structures that operate in the mind of the listener.

In contrast to language which has an undeniable survival value, music is still controversial in this respect [Pinker, 1997]. Contrastingly too, music is predominantly self-referential instead of referring to a world that is outside its own symbols, as it is the case of language [Besson and Friederici, 2005]. Despite these differences, music exerts powerful effects on individuals, couples, families, work groups, tribes, and also on entire societies [Hargreaves and North, 1997]. In order to clarify the questions posed by music, empirical methodologies are required. Experimental data from psychoacoustics and music psychology exists since the middle of the XIXth century only [Helmholtz, 1863], and has led to the specification of the thresholds and resolution of duration, pitch, loudness, and timbre (and their respective interactions), in addition to rough characterizations of the different memory systems of the brain and the processes of encoding and decoding information in and from these systems. This empirical approach has benefited from recent techniques of brain imaging, which provide hints to modularization, hierarchy, and interdependence of the cognitive architecture's elements. As we will see, the current trends favor synergetic approaches combining knowledge coming from these disparate sources into computational models that help to explain existing data but also predict yet unobserved phenomena and results.

The pages that follow have been organized as follows. We first describe the methodology employed to retrieve insight into the phenomena of music perception and cognition. Then we show how brain measurement hints towards a particular architectural organization of music cognition. We follow the chain of perceptual and cognitive processes that operate when we listen to music: We discuss findings from neuroscience and their implications for musical feature extraction. Then we summarize psycho-acoustic and psychological research

relevant for perception, before addressing high level functions such as different kinds of memories, attention, and expectation. But before summarizing the findings on which music cognition models can be based, we describe the (experimental) methods that yield them.

2 Methodology

The relevant methodology ranges from invasive and non-invasive neurophysiological experiments in animals and humans to psychoacoustic and music psychology experiments as well as music theory.

2.1 Brain Measurement

Invasive recordings of neural activity in animals are the principle source of knowledge about neural correlates of hearing. [2] Can such recordings in animals also be employed for the investigation of music cognition? The relevance of animal experiments for music cognition research depends on how the musical abilities of the research animals and humans compare.

How far can brain imaging in humans help the understanding of music cognition? Sometimes diagnosis of epilepsy in humans indicates the need of electrode implantation in the patient for localizing the center of the epileptic seizure. This may give the opportunity of invasive music cognition experiments. We can conceive the brain as a net of approximately 10^{11} neurons (Kandel et al. [1991], p. 18), each often having a complex topology and synapticly connecting to up to as much as 150 000 other neurons. [3] Hence, in general it is impossible to localize electric current flow in individual neurons by measuring electricity in electrodes positioned around the skull, like in the electroencephalogram (EEG). Event-related potentials (EVP), voltage deflections following a previous (unpredictable) stimulus, reveal insight into expectedness of a sound [Koelsch and Siebel, 2005]. Mismatch negativity (MNN) is a negative potential that is related to infrequent changes in repetitive acoustic sequences [Näätänen et al., 2001]. It stems from the auditory cortex and reflects regularities in the acoustic environment. Similar to the EEG, the magnetoencephalogram (MEG) detects neuro-magnetic brain signals. Other non-invasive methods, such as functional magnetic resonance imaging (fMRI) or positron emission

[2] Recently, technologies such as the two-photon laser scanning microscope [Majewska et al., 2006] for in vivo imaging provide the technical basis for further progress in the field.

[3] Purkinje cell in the cerebellum (Kandel et al. [1991], p. 22).

tomography (PET), have better spatial but worse temporal resolution. These methods are especially applied in patients with brain lesions in conjunction with psychological experiments and its application in music research is still an exploratory domain.

2.2 Psychology and Musicology

Most musical qualities correspond to *psychoacoustic* qualities rather than physical ones. In order to learn about such qualities various experimental methods are used, such as verbal response, response time and accuracy, as well as (for babies) head turning, sucking strength, and heart-rate deceleration [Dowling, 1999]. As Chomsky [1988] claims for language, the deep structure of music cannot be a formalized rule system extracted solely from a corpus of music. The major reference is the musician or naive music listener, assessed as subjects in a psychological experiment. But also in music psychology the surface level could be falsely taken for the deep structure, because of two reasons: 1) The "role that anthropological field-workers [musicologists in our context] play in shaping, rather than merely describing, the situations they report on" [Britannica, 2003]. E.g. music students are educated using musicologist's concepts. In some cases, such as music psychology experiments with probe tones or circular pitch [Purwins, 2005], the judgments of musically educated subjects reflect the application of explicit music theoretic concepts rather than some deep structure inherent in the music itself. 2) In our global musical monoculture, it is difficult to find unbiased naive subjects that have never heard mainstream music before. So far, music psychology has predominantly studied simple musical attributes such as pitch, tempo, or rhythm. Due to the lack of empirical data from music psychology, for more complex musical attributes we need to refer to traditional music theory.

2.3 Models and Implementations

What is a cognitively relevant *representation* of music? In music theory, categories beyond score notation have been searched for. Theories of harmony provide examples how to subsume chords in different harmonic categories, e.g. functional harmony assumes three classes of harmony: tonic, dominant, and subdominant [Riemann, 1877]. Schenker [1935] extracts abstractions from a score, step by step reaching a more compressed representation on temporal layers of coarser temporal resolution. The limitations of score notation as a correlate of "musical objects" become apparent. Indications of this are the constructs of "virtual" notes that do not appear in the score, e.g. the usage of the dominant chord with omitted fundamental [de la Motte, 1980] or notes in

the middle ground level analysis not existent in the score in Schenkerian analysis [Schenker, 1935]. In the 90s, the machine listening group at MIT searched for cognitively plausible "musical objects" [Ellis, 1996; Scheirer, 2000], but it turned out that there is still a long way to go to identify which these objects might be.

In chunk formation, such as event fusion, grouping, and segregation on the level of musical form, different cues compete, like pitch and temporal proximity for tones that are close in pitch but distant with respect to their onsets. Based on laboratory experiments with simplified stimuli, the rather unspecific character of the *gestalt* principles suggested by Bregman [1990] poses difficulties to their practical application.

How fine-grained can we differentiate within one musical feature? Quantization is well studied for pitch and loudness, e.g. by investigation of *thresholds* and *just noticeable differences*. The differentiation and borders of categories in the domains of rhythm and harmonies are of great interest but less understood than discrimination of basic features such as pitch and loudness.

Music has been compared to *language* as well as to *motion* processes. The aspects it shares with language have been described with methods from artificial intelligence, e.g. finite state grammars and graphs. But other aspects of music can better be described as processes. [4] To account for *adaptation* and *learning*, artificial neural networks have been used. Artificial neural networks have been advertised for their higher biological plausibility, which is sometimes equivocal [Purwins, 2005]. However beyond the biological allusion of the terminology, problems of statistical machine learning and numerical mathematics, e.g. for the solution of differential equations, have to be solved.

Efficient modeling of aspects of human *memory* is a problem addressed in various ways: 1) transition tables - usually of first order - used e. g. in Bayesian networks such as hidden Markov models or Kalman filters, 2) a leaky integrator modeling echoic memory, 3) a hierarchical model that works analogously on each temporal resolution level [Todd, 1991; Mozer and Das, 1992], 4) long short term memory [Hochreiter and Schmidhuber, 1997], and 5) a toolkit such as ACT-R [Anderson et al., 2004] that are designed to explicitly model chunking mechanisms in memory.

2.3.0.1 What is a Valuable Cognitive Model? A model is evaluated by its predictive and generalization power, its simplicity and its relation to existing theories. The model *predicts* behavior that is later displayed in an experiment. The behavior would be unexpected if the model would not have been taken into account. A model should *generalize*. It should be able to

[4] Lidov [1999] tracing back to Schopenhauer [1859], Nietzsche, and Herbart.

explain data (e.g. results of experiments) that have not been taken into consideration in building the model. Generalization capability is supported if the *complexity* of the model (e.g. number of parameters, complexity of formulas) is small relative to the amount of data it explains. Otherwise overfitting may occur [Vapnik, 1998]. Another feature of a model is to what degree it shows a relation between two formerly unrelated theories. It may *unify* existing theories. This may help understanding a phenomenon instead of just reproducing input-output relation inherent in the data. Another aspect is the quantity and diversity of empirical data the model explains. E.g., there are models that are only based on a single empirical measurement. However, if the model is built to fit all available data this may not imply that the model can generalize well. Of course, in the model, *assumptions* and consequences must be clearly defined and distinguished.

We first discuss the impact of computational neuroscience for music cognition models. Then we give an overview and criticism of music models for perception, grouping, and high level cognition.

3 Perceptual and Cognitive Processing Chain

3.1 Modular Architecture of Music Processing

Can we infer information about the pathway of music processing from brain measurement? Are there parallel processes? Is information processed hierarchically? Experimental work in psychology, electrophysiology, brain imaging, and lesion studies support the hypothesis of a modular organization of music processing (summarized in Peretz and Zatorre [2005], cf. Table 3.1).

The pathway of auditory processing is roughly as follows: After being filtered by the outer/middle ear, the cochlea and the hair cell transform the sound signal into electrical activity further propagated through the superior olivary nuclei, the inferior colliculus, the medial geniculate nuclei (thalamus) and the primary auditory cortex. The auditory cortices are nested inside each other with the primary cortex forming the kernel and the tertiary auditory cortex being the outermost one. Whereas basic musical features such as pitch and loudness are processed by the primary auditory cortex, the "secondary cortex is believed to focus on harmonic, melodic, and rhythmic patterns and the tertiary auditory cortex is thought to integrate these patterns into an overall perception of music" [Abbott, 2002]. The auditory cortex closely interacts with Broca's (grammar processing) and Wernicke's (perception of speech prosody) areas and the motor cortex. From neural response times, measured as event related potentials, Koelsch and Siebel [2005] derive the following music pro-

cessing chain: feature extraction, gestalt formation, interval analysis, structure building, structural reanalysis and repair, vitalization, and (premotor) action. They indicate a feedback relation from structure building to feature extraction.

A distinction between *pitch* and *time-based* relations is supported by the majority of the experiments, even if there is no clear boundary in the brain map [Justus and Bharucha, 2001; Krumhansl, 2000]. Pitch intervals and contours (the ups and downs of pitch sequences) are processed in different regions than absolute pitch [Zatorre, 1985; Ayotte et al., 2000], as are complex tones [Hall et al., 2002; Hart et al., 2003]. Some rather abstract tonal relations have been robustly located in the brain (harmonic violations, Regnault et al. [2001]; consonance/dissonance, Fishman et al. [2001]), while others have not yet been precisely located (scale, chords).

Basic time relations can be divided into two processes: meter and grouping. Hereby, meter refers to the detection of a regular beat (or pulse) that leads to metrical organization based on periodic alterations between strong and weak beats. "Grouping refers to the segmentation of an ongoing sequence into temporal groups of events based on their durational values"[Peretz and Zatorre, 2005]. Meter and grouping seem to operate independently rather than hierarchically. Experiments indicate that the right hemisphere better handles meter formation [Fries and Swihart, 1990; Wilson et al., 2002] whereas the left hemisphere accounts for grouping. In addition, cerebellum and/or basal ganglia possibly are responsible for motor and perceptual timing. Motor and premotor areas seem to be involved in this task (Ivry and Keele [1989], Penhune and Doyon [2002],Janata and Grafton [2003]). Tramo [2001] mentions distinct patterns of activation elicited by the processing of musical material with simple and complex meter. We will now specifically look at models of the auditory periphery.

3.2 Auditory Periphery

Lyon and Shamma [1996], p. 227, give a general characterization of auditory models [5]:

> All models, however, can be reduced to three stages: analysis, transduction, and reduction. In the first stage, the unidimensional sound signal is transformed in a distributed representation along the length of the cochlea. This representation is then converted in the second stage into a pattern of electrical activity on thousands of auditory nerve fibers. Finally, perceptual

[5] In this section we partly follow Normann [2000].

Ability		Localization	Reference
Pitch-based	Tones (complex)	posterior sec. cortex	Hall et al. [2002]; Hart et al. [2003]
	Pitch (f0)	posterior sec. cortex	Zatorre [1988]; Liégeois-Chauvel et al. [1998]
	Chroma	anterior sec. cortex	Warren et al. [2003]
	Intervals	r. temp. cortex	Zatorre [1985]
	Contours	r. superior temp. gyrus	Ayotte et al. [2000]; Peretz [1990]
	Harmony	inferior frontal area, Heschl gyrus (cons/dis)	Regnault et al. [2001]; Fishman et al. [2001]
	Melody Categorization	l. inferior frontal and superior temp. region	Ayotte et al. [2000]; Platel et al. [1997, 2003]
Time-based	Grouping	l. hemisphere	Vignolo [2003]; Pietro et al. [2004]
	Meter	r. temp. aud. cortex, superior temp. gyrus	Fries and Swihart [1990]; Wilson et al. [2002]
	Meter (simple)	l. frontal, parietal cortex, r. cerebellum	Tramo [2001]
	Meter (complex)	r. frontal shift	Tramo [2001]
Motor/ perceptual timing		basal ganglia, cerebellum	Ivry and Keele [1989]; Penhune and Doyon [2002]; Janata and Grafton [2003]

Table 1
Summary of localization of musical activities in the brain, from Peretz and Zatorre [2005]. We use the following abbreviations: r.=right, l.=left, temp.=temporal, aud.=auditory, sec.=secondary, cons/dis=consonance/dissonance.

representations of timbre and pitch are extracted from these patterns in the third stage.

In this three-level description, analysis refers to linear processing. Transduction and reduction are non-linear processes.

With reference to biology, the linear part concerns the *outer* and *middle ear*, and the basilar membrane. Transduction refers to the non-linear transformation by the inner hair cell, and, in a broader sense, the stochastic transmission at the auditory nerve fiber. Finally, the "mechanical vibrations along

the *basilar membrane* are transduced into electrical activity along a dense, topographically ordered array of auditory nerve fibers" [Lyon and Shamma, 1996].

Both outer and middle ear combined show a response curve maximal between 1 and 3 kHz and decreasing towards lower and higher frequencies.[6] For imitating various pitch phenomena the effect can be implemented as a band pass filter.[7] The inner ear contains the basilar membrane, with hair cells adjoined to it. Each frequency of an incoming sound sets the basilar maximally into motion at a particular position. The logarithm of the sound frequency and the distance between the position of maximal motion and the apex of the basilar membrane are approximately proportional [Kandel et al., 1991]. This mapping is related to phenomena such as roughness, masking (when one sine tone is made inaudible by another louder one with proximate frequency), and just noticeable frequency differences.

The major properties of the *hair cell* in the inner ear are temporal coding and adaptation behavior. According to Meddis and Hewitt [1991], the synapse of the hair cell can be described as a dynamic system consisting of four elements. This hair cell model reproduces a limited number of experimental data from hair cells of gerbils, in particular: (i) Frequent spikes occur with the onset of a tone. The spike rate decreases to a constant value, when the tone continues, thereby revealing adaptation behavior. After the offset of the tone, it decreases to about zero. (ii) Below 4-5 kHz spikes occur in the hair cell nearly exclusively during the positive phase of the signal (*phase locking*). In this range, thus, frequency is coded both by the position of the responding hair cell on the basilar membrane and by temporal spiking behavior. For frequencies above 5 kHz, spikes occur about equally often during the positive and the negative phase of the signal. Therefore, above 5 kHz, frequency is only coded by the place information on the basilar membrane.

Auditory processing in the auditory periphery including the hair cell is well known and successfully applied in music compression, sound analysis, and music cognition modeling [Leman, 1995; Rosenthal and Okuno, 1998]. Contrastingly, the physiological correlate of the reduction stage is understood very poorly. Auditory models are quite speculative at that stage. In using embedding and clustering algorithms, we can claim biological plausibility only within narrow bounds, e.g. we can mimic the auditory principle of tonotopy. For simulation environments of early processing confer Appendix A.

---

[6] Meddis and Hewitt [1991], p. 2868, Figure 2. Similar filter curves are motivated by different psychoacoustic phenomena, such as hearing threshold [Yost and Hill, 1979] or dominance region [Terhardt et al., 1982].

[7] IIR filter, Oppenheim and Schafer [1989], of second order, Meddis and Hewitt [1991], footnote p. 2881.

3.3 Some Principles of Neural Processing

In computational neuroscience there are various theories about how information is coded and how neuronal dynamics can be numerically simulated. In neurobiology, Hebbian learning (Section 3.3.2) is a basic principle that is also related to proximity preserving brain mappings from sensory input to brain areas such as the cortex.

3.3.1 Information Coding Hypotheses

There are three major hypotheses which attempt to explain how a sequence of spikes encodes information: (i) by exact neuron firing times, (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 non-linear differential equations [Hodgkin and Huxley, 1952] describing current flow in the axon. As a simplification, we can make use of the "integrate-and-fire" model introduced by Peskin [1975]. [8] Voltage is integrated until threshold is reached. After a refractory period, integration starts again from rest potential. (iii) is a rough simplification of spike behavior. Nonetheless, (iii) is the basis of the connectionist neuron used in multilayer feed forward networks in artificial neural networks (cf. e.g. Haykin [1999], p. 156 f.). Neuron [Carnevale and Hines, 2006] is a popular framework for low-level neural dynamics simulations.

3.3.2 Hebbian Learning and Tonotopy

If presynaptic and postsynaptic electrical activity occur synchronously, the postsynaptic receptor channels become more permeable so that a presynaptic activity evokes stronger activity at the postsynaptic dendrite (*Hebbian learning*).

According to the principle of *tonotopy*, proximate hair cells on the basilar membrane project to proximate neurons in the brain. In computer science, the tonotopic principle can be implemented by algorithms such as the self-organizing feature map [Kohonen, 1982].

[8] See Maass [1997] for a computational analysis.

3.4 Discrimination

We are interested in the minimal detectable value (threshold) and the just noticeable difference (JND) of sound attributes, e. g. pitch or loudness [Roederer, 1995]. According to the Weber-Fechner law, a multiplication of stimulus intensities corresponds to an addition of the perceived stimuli. The law assumes that the JNDs are additive. If we would know the JNDs for all perceptual dimensions we would have a discretized grid of the multidimensional sensory space, i.e. we would have the " 'stones' on which the 'house of sensations' is built" (Zwicker and Fastl [1999], p. 175). The Weber-Fechner law is implemented in the sound pressure level (as a measure of the perceived quality, measured in dB) calculated from the sound pressure. The JND for the sound pressure level is about 1 dB (Zwicker and Fastl [1999], p. 175). We can quantify the perceived difference between two stimulus intensities. A doubling of the perceived loudness roughly corresponds to an increase of the sound pressure level by 10 dB.

The results of studies on low-level features indicate psychophysical limits in our ability to perceive duration and durational succession. The minimal inter-onset interval to perceive two onsets distinctly is 2 ms. At least 10-15 ms are necessary to determine the order of the onsets [Hirsh, 1959]. Reliable judgments of length require a minimal inter-onset interval of 100 ms [Hirsh et al., 1990]. This can be related to the processing time in the cortex [Roederer, 1995]. [9]

3.4.1 Pitch

Even though for an acoustical signal, frequency and pitch are usually closely related, it is not generally possible to reduce pitch perception to the physically measurable attribute of frequency. As an example, we will discuss the phenomenon of missing fundamental. Neglecting transient and noise components, if they exist, we can treat tones as approximately periodic signals, as a rough approximation of the singing voice.

There are three different approaches for pitch detection:

- In a spectral model, the pattern of spectral peaks of the signal is analyzed. As an example, let us consider tones composed of sine tones with fundamental frequency f_0 and integer *multiples* $f_k = k \cdot f_0$ (the *partials*) of the fundamental frequency. If we remove the sine with fundamental frequency we still hear a tone with a pitch corresponding to that frequency. As an

[9] For the neural foundations of early processing cf. Kandel et al. [1991]. For loudness and pitch perception, including masking, see Zwicker and Fastl [1999].

example of a spectral method we will explain Terhardt [1974]'s approach to the phenomenon of missing fundamental. We first need to introduce the *subharmonic* f'_k which is an integer *fraction* $f'_k = f_0/k$ of the fundamental frequency f_0 (cf. Figure A.1). Terhardt [1974] considers subharmonics of all partials within a certain frequency range. The perceived pitch is the pitch that can be considered a subharmonic of most partials.

- In a temporal model, the regularities in the temporal domain are detected, e.g. by autocorrelation. Pitch detection is often implemented by autocorrelation that detects periodicities in the temporal domain. But biological evidence for a process like autocorrelation is doubted [Meddis and Hewitt, 1991]. Temporal and spatial models are closely related [de Cheveigné, 2005].
- In a spectrotemporal approach, first the signal is filtered by a model of the cochlea and then a temporal model is applied. Spectrotemporal methods combine a model of the basilar membrane with a subsequent temporal method such as autocorrelation. Shamma [2001] argues that pitch perception can be modeled without the implausible need of "specialized neural machinery for analyzing temporal input" (oscillators, delay lines). He suggests that a "unified computational framework" is used for both visual and auditory processing. In analogy to receptive fields in the primary visual cortex, spectrotemporal response fields in the primary auditory cortex are proposed that let harmonic templates emerge.

3.5 Grouping

Certain sensory features, such as detected partials, may be grouped together, forming an entity (*gestalt*), such as a tone, according to particular criteria. The gestalt concept originated from Ehrenfels [1890] and Mach [1886] and was established later by Max Wertheimer, Wolfgang Köhler, and Kurt Koffka. Initially musical examples were presented. Subsequently, visual perception was investigated. From the 1970s onward, computer-supported sound synthesis and analysis enforced the application of gestalt theory to auditory perception, exhaustively reviewed in Bregman [1990].

In the following, principles are introduced which aid grouping in auditory perception (Figure A.2, Bregman [1990]): The principle of *proximity* refers to distances between auditory features, e.g. 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 A.2 A 1) which contains large pitch jumps, is perceived as one stream. The same sequence of notes played very fast (Figure A.2 A 2) produces two perceptual streams, one stream consisting of the As and another one consisting of the Bs.

Similarity is very similar to proximity, but refers to properties of a sound, which cannot be easily identified along a single dimension (Bregman [1990], p. 198). For example, we speak of similar rather than of proximate timbres.

The principle of *good continuation* denotes 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 A.2 B) high (H) and low (L) tones alternate. If the notes are connected by frequency glides (Figure A.2 B 1) both tones are grouped to a single stream. If high and low notes remain unconnected (Figure A.2 B 2) the Hs and the Ts form separate streams. "Good continuation" is the continuous limit of "proximity".

The principle of *closure* completes fragmented features, which already have a "good gestalt", e.g. ascending and descending frequency glides that are interrupted by rests in a way depicted in Figure A.2 C 2. Separated by rests, three frequency glides are heard one after the other. Then noise is added during the rests, as shown in Figure A.2 C 1. This noise is so loud that it would mask the glide, even if it were to continue without interruption. Surprisingly, the interrupted glides are perceived as being continuous. They have "good gestalt": They are proximate in frequency and glide direction 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 or as an active perception that allows filling perceptual gaps.

The principle of *common fate* groups frequency components together, when similar changes occur synchronously, e.g. onsets, glides, or vibrato. Chowning [1980] (Figure A.2 D) made the following experiment: First three sine 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 tone 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, as each one seems to have a "different fate".

In the evolution of the auditory system, grouping principles serve as a means of identification of sound sources. But in music there is not a one-to-one correspondence between tones and sound sources. A single instrument can play several tones simultaneously. On the other hand, one tone can be produced by several instruments in chordal streams. This particular challenge to the auditory grouping principles could partially explain our fascination for music.

3.5.1 Implementation of Event Fusion and Grouping

We first introduce two hypotheses on the mechanism underlying event fusion (Section 3.4) as well as grouping (Section 3.5), then describe their implementation, cover computational models of streaming and their application to the related problem of acoustical source separation.

3.5.1.1 Hypotheses on Grouping

Two hypotheses regarding the neural implementation of grouping, more generally speaking, the binding problem, are given (cf. Treisman [1999]).

The first hypothetical solution to grouping, *hierarchical organization*, 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 [Semple and Scott, 2003]. For example, they respond to partials, tones, chords, harmonies, and keys. This corresponds to hierarchical artificial intelligence approaches (cf. context-free grammars). Even though structures in the brain include lateral connections (breaking the hierarchical organization), in practice a strictly hierarchical concept is successful, e.g. within a framework of a knowledge database and a Bayesian network [Kashino et al., 1998].

Another way of trying to explain event fusion and grouping is through *neural synchronization*. 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 that 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 synchronization in the visual cortex is given by Engel et al. [1997], and in the auditory cortex by Eggermont [2000]. However, Shadlen and Movshon [1999] argue that temporal signals do not have the capability of encoding binding relations [Roskies, 1999]. We will now look at a specific implementation of grouping through neural synchronization.

3.5.1.2 Oscillator Model for Grouping

Terman and Wang [1995]; Wang [1996]; Brown and Cooke [1998]; Wrigley and Brown [2005] supply an implementation of event fusion and grouping through neural synchronization. Wang

[1996] introduces a two-dimensional grid of oscillators.[10] The oscillators are arranged according to frequency in one dimension and a time delay line in the other dimension. The closer the oscillators are the more they are coupled, thereby reflecting Hebbian learning (Section 3.3.1) and the principle of proximity (Section 3.5). An additional global inhibitor prevents oscillators belonging to different streams from being active at the same time. Stream segregation of visual and musical stimuli in symbolic representation are presented. Brown and Cooke [1998] use another oscillator grid after preprocessing music input by a gammatone filter bank, a hair cell model, and an onset map. Applications are shown for segregation of vowels and for different voices [Cooke and Brown, 1999]. An approach closer to the biological reality would be based on the "integrate and fire" model [Maass, 1997]. An ensemble of such models displays synchronous spike patterns.[11]

3.5.1.3 CASA In *computational auditory scene analysis* (CASA), grouping principles are translated into numerical algorithms. Two questions are of particular importance: Which is the domain in which the principles are applied (e.g. spectral domain)? What is a meaningful representation? Provided that an appropriate representation is given, how can we translate the principles into numerical equations? How to design an appropriate distance measure, e.g. for proximity? CASA can play a role in a prominent engineering problem: the separation of sound sources.

3.5.1.4 Acoustical Source Separation A mixture of several sound sources (speech, music, other sounds) is recorded by one or several microphones.[12] The aim is the decomposition into the original sources. A particular goal is demixing with at most two microphones.

The approaches can be roughly divided into two categories:

(1) Mimicking the auditory system and
(2) Employment of techniques of digital signal processing without reference to biology.

Okuno et al. [1999] aim to combine (1) and (2) synergetically. A possible treatment similar to (1) is as follows: An auditory model is used for preprocessing. From the output of the auditory model (cochleagrams and correlograms) harmonic substructures are extracted. By the use of grouping principles, spectral

[10] For a *Matlab* implementation see the *Matlab* Audio Demonstrations by Cooke et al. [1997].
[11] Personal communication with Gregor Wenning.
[12] This paragraph is taken from Purwins et al. [2000a] in a slightly adapted form.

units are built from this. From these separated units, sound can be resynthesized [Nakatani et al., 1995]. In another approach [Slaney, 1994, 1998b], the spectral units are determined as a sequence of correlograms, and the auditory transformation is inversely calculated in order to resynthesize the original signal from its encoding.

The larger portion of methods according to (2) deals with the realization of the *independent component analysis* (ICA, Comon [1994]; Cardoso [1998]; Müller et al. [1999]). Although Sanger [1989] indicates a reference to biological systems, his approach is largely a statistical model. Decomposition of mixtures in a real acoustic environment works under very special conditions only. Approaches in this direction are Lee et al. [2000]; Casey and Westner [2000]; Parra and Spence [2000]; Murata et al. [2001].

3.6 Short Term Memory and Attention

Other important topics in auditory perception are *attention* and *memory encoding*. 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 steady, then that aspect is drawn to the listener's attention (*figure ground segregation phenomenon*). Let us give an example for memory encoding: The perceived illusory continuity (cf. Figure A.2 C) of a tune through an interrupting noise is even stronger, when the tune is more familiar (Bregman [1990]: p. 401).

Deliège [1987, 1989] hypothesizes the operation of a cue-abstraction mechanism. The cue would be a "very restricted entity often shorter than the group itself, but always embodying striking attributes" (Deliège [1989]: p. 307) which suffice to signify longer entities or groups. The cue abstraction mechanism may, in this sense, be compared to the figure ground segregation phenomenon. Information is reduced by focusing attention on small features that can be distinguished from a more diffuse or global environment. The cues play the figural role and become abstractions used to lighten the load on memory storage. It should be noted that, as in the figure ground segregation phenomenon, the perceptual qualities of the cue are differentiable from the environment from which it is abstracted. As it is the case for a figure, a cue would be more salient and consequently more precisely defined than elements of its environment (or "ground") which would be less structured [Deliège, 1992].

Memory limits also need to be considered in the encoding of musical material. A musical pattern must satisfy a number of constraints in order to be easily encoded and then to contribute to the perception of connectedness between groups of musical events separated in time [McAdams, 1989]. The event has to

be short enough (about 2-5 seconds; Fraisse [1957]); it has to be unified enough to be grouped into a chunk in short-term memory (according to Miller [1956], a limit on short-term storage is classically set at about 5-9 chunks, though this depends on the nature of the pattern); patterns that are organized according to easily discernible rules of construction, such as hierarchical patterns, are generally remembered more easily and can have more elements in short-term memory than patterns organized in other ways [Deutsch, 1999]. Memorability may also be greater for patterns that have some kind of inherent stability or well-formedness as well as for patterns that are more easily assimilated to an existing knowledge structure. What "well-formedness" means may be quite different for different dimensions and it is not clear to what extent it would be independent of cultural convention, and thus of acquired knowledge structures.

3.6.0.5 Analytic and Synthetic Listening Attention and learning favor either of two opposing auditory strategies. In *analytic listening* particular components, e.g. partials of a sound or notes of a melody, are singled out. *Synthetic listening* gives rise to holistic perception, e.g. a tone is perceived as a whole, indivisible entity, a melody is perceived as a single object (a profile), instead as a sequence of discrete notes. In musical composition often both strategies are supported. For instance, in Webern's orchestration of the *Ricercar* in Bach's *Das Musikalische Opfer* a competition of grouping principles takes place in the domains pitch and timbre. Focusing on pitch as a cue, the melody is heard. Focusing on timbre the main line decomposes into small phrases. Bregman [1990] (p. 470) points out the visual analog in Giuseppe Arcimboldo's paintings. Analytic listening seems to be a top-down kind of activity (e.g., attention is guided by existing knowledge), whereas synthetic listening is a bottom-up activity (e.g., attention is guided by those stimulus features that are salient). These different processing modes influence the type of encoding and, hence, the durability, interference and usability of the representation that has been stored: analytic processing paves the way for conceptual learning, whereas synthetic processing is related to perceptual learning.

3.6.1 A Computational Model of Attention

Wrigley and Brown [2005], Section 3.5.1.2, extend Wang [1996]'s oscillator grid. They refine the system by adding cross-channel correlation information (in order to form segments) and use an attentional leaky integrator (ALI) to form an attentional stream. Acoustic features are "attended" if their oscillatory activity is synchronized with a peak in the ALI activity. The system is evaluated by reproducing three perception experiments with tones composed of sinusoids with a frequency that equals an integer multiple of the fundamental frequency of the tone: First, the segregation of a detuned partial [Darwin

et al., 1995] is considered. In accordance to experiments, in the model, in case of a just slightly detuned fourth partial, all sine components fuse to one tone. If the detuning exceeds a critical limit, the stimulus segregates into two tones, a singled-out partial and the rest. Second, the model reacts according to the old-plus-new heuristics in which the non-detuned fourth partial is presented alone before the other sinusoidal components join in [Darwin and Ciocca, 1992]: The fourth partial singles out even if it is in tune. Third, the model is consistent with a binaural two-tone streaming experiment with an attentional distractor on one ear [Carlyon et al., 2001]. However, the model has been validated only with these rather artificial stimuli.

3.7 Long Term Memory, Schemata, and Expectation

Whereas innate auditory grouping principles perform low level processing, the learning of categories and schemata is part of high level processing. Two different types of musical knowledge are invoked as constructs guiding explanations that require access to long-term memory: *categories* and *schemata*. Categories are conceptual devices that help to create bundles of musical elements somehow associated in some context. When several musical objects (e.g. notes, chords, phrases, articulations, etc.) share a set of properties or present similarities, they are candidates to be assigned to the same category (e.g. "C♯", "g–minor", "Phrygian scale", "appoggiatura"). Without the help of categories, two musical notes with frequencies differing by a few cents could not be assigned to the same class. All the small nuances of musical stimuli would create a cumbersome and difficult to understand experience. Just noticeable difference and thresholds determine the quantization of pitch and loudness (Section 3.4). In a similar way, categories perform discretization on a higher, later level of processing, resulting in discrete elements that can better be managed by our memory systems [Snyder, 2000]. Although most of the categories used by humans are cultural constructs, and hence, musical categories are too, there is some debate on the possibility that certain categorical distinctions can be universal (e.g.,"living thing" versus "inanimate object", Farah and Rabinowitz [2003]). What is clear is that the cognitive system is very flexible in the sense that it allows different systems of categories to develop in order to represent the perceptual environment, depending on the statistical structure of the environment [Barlow, 1989] and on the education that every human being receives.

Categories are somehow static constructs, and when the temporal dimension needs to be included in a representation we need to invoke schemata. Schemata help to recall and recognize series or combinations of musical objects and events. They also guide our attention and facilitate the elaboration of predictions about the evolution of musical events. The idea of the existence of

schemata dates back to Kant. During the first half of the twentieth century the psychologist Bartlett popularized the term through an influential definition of schemata as involving a series of organized responses corresponding to a series of stimuli [Bartlett, 1932]. A usable definition of the contemporary meaning is offered by Mandler, who describes a schema as a knowledge structure "formed on the basis of past experience with objects, scenes, or events and consisting of a set of (usually unconscious) expectations about what things look like and/or the order in which they occur" [Mandler, 1979]. Gjerdingen [1988] points out that these expectations are both activated by and concerned with higher-level events. Bregman [1990] (p. 734) emphasizes: "In all cases it is abstract enough to be able to fit a range of environmental situations." Schemata can represent information of various types on different levels and can be embedded, one schema containing other ones [Rumelhart, 1980]. Tonal schemata, for example, are formed through repetitive experience with melodic material and are a type of long term memory representation involving contour and pitch class. From a computational point of view, schemata have been implemented as frames, scripts or conditional rules [Holland et al., 1986]. Leman [1995]; Purwins et al. [2000b] have worked with self-organizing maps as a scaffolding for the development of musical schemata.

Categories and schemata can be learned by means of supervised learning mechanisms (i.e. having a teacher who provides appropriate training examples and exercises, like in formal music education). However, in general, learning occurs in an unsupervised manner, i.e., by being exposed to a given set of structured stimuli that obey some regularities [Tillmann et al., 2001].

3.7.1 High Level Cognition Models

In this subsection, we give an overview of computational models of higher level cognitive processes. We review some methods for category formation, a central topic in cognitive modeling. Then we present ACT-R, a general computational framework for cognition. Lastly, we describe two specific cognitive processes that are closely related: musical sequence learning and expectation in music listening.

3.7.1.1 Category Formation An important aspect of cognition is the formation of perceptual categories. Several computational approaches exist to address this task, with different points of focus and different degrees of cognitive relevance.

The *self-organizing feature map* (SOM, Kohonen [1982]) is frequently presented as a model of cortical organization [Obermayer et al., 1990]. Its backbone is the winner-takes-all step, requiring a global comparison between all

neurons. Due to the high number of neurons in the cortex, such a step is biologically implausible. Although we know little about the cortex, the SOM is used as a rough implementation of instances of the schema concept [Leman, 1995; Purwins et al., 2008] or categorization processes.

The underlying principle of the *adaptive resonance theory* (ART, Grossberg [1976]) architecture is the trade-off between plasticity and stability in category formation. We illustrate this in the context of a classification task: when processing an input vector, the model looks for a similar stored pattern within a certain range of tolerance. If matching patterns are found, the most similar stored pattern is modified to incorporate the input vector. If no pattern matches the input, a new category is created by storing a pattern that resembles the input. Stability refers to the former process: the ability of the system to robustly interpret its environment in terms of the internal patterns it already has. Plasticity refers to the latter process: the ability to adapt to changes in the environment by the formation of new categories.

According to De Barreto et al. [2003], despite some similarities, the SOM and ART networks differ in the way they categorize input data. Indeed, they have been designed to tackle different aspects of learning. The SOM is concerned with the preservation of neighborhood relations: nearby output neurons should store nearby input patterns. The ART network, in turn, is concerned with the above mentioned stability-plasticity dilemma.

Another approach to unsupervised categorization of data is based on the process of concept formation. COBWEB, by Fisher [1987], is a model of data driven taxonomy emergence. The model is based on the maximization of a heuristic function by applying merging and splitting operations to the hierarchical partition of the input space. The heuristic function (*category utility*) is based on the attribute prediction accuracy given a partition. Is has recently been used by Marxer et al. [2007] to model the emergence of scales, motivic, and harmonic categories.

3.7.1.2 Integrated Cognition Modeling Environments It has been shown that listening to piano music evokes (pre)motor activity in the brains of pianists [Haueisen and Knoesche, 2001]. Such findings show that action planning for movement coordination is involved in music perception. Integrated cognition modeling environments provide a platform to model such interactions between action and perception.

Adaptive control of thought (ACT-R, Anderson et al. [2004]) is an architecture that unifies modules for various cognitive functions: the perceptual-motor module, the goal module, the declarative memory, and the procedural memory module. ACT-R as an architecture can be used to implement models for a

variety of cognitive tasks, such as language processing, game playing, mathematical problem solving, and car driving. The task execution is performed in a goal-driven manner, where the system tries to reach a given goal by deriving subordinate goals through the production rules and subsequently pursuing these. The latency and accuracy of the model's performance is proportional to the the number of subgoals. Through a spreading activation mechanism, chunks are activated, and based on the activation patterns of the chunks, the best production rule is selected competitively. Chunk activation and connexion strengths are learned in a Bayesian manner. A prior is given to chunks and production rules, and from several instances of observations the learned probability for a task is updated. New production rules can also be generated based on old rules (production compilation). Figure A.3 shows a diagram of the ACT-R architecture.

ACT-R simulations can be evaluated by matching the model's performance against human's performance time and accuracy of the same task. Response times can be compared, since the model assigns a latency to the execution of each step in the various modules. In addition, ACT-R can be evaluated by linking ACT-R's modules to specific brain regions. Consequently, a task simulation then predicts a particular flow of activation through the brain. The predicted flow can be compared to brain activity in humans that execute the task, observed in a brain imaging experiment such as fMRI. Prediction and measurement of latency, accuracy, and activation-flow is not limited to skilled task-execution. Especially comparing prediction and measurement during the learning phase of the task can give a good insight in the validity of the model.

3.7.1.3 Information Theory and Expectancy Abdallah and Plumbley [2007] state that "perceptible qualities and subjective states like uncertainty, surprise, complexity, tension, and interestingness are closely related to information-theoretic quantities like entropy, relative entropy, and mutual information." In Pearce and Wiggins [2004], *entropy* and *cross-entropy* are used to evaluate whether statistical models can learn symbolic monophonic melodies. The *entropy rate*, denoted $H(X|Z)$, reflects the instantaneous certainty of statistical models of characterizing the current observations X given past observations Z, while *cross-entropy* applied to test melodies informs about the generalization accuracy of a learned statistical model. Abdallah and Plumbley [2007] extend these ideas and propose to compute the *average predictive information rate*, noted $I(X, Y|Z)$ which may be seen as the average rate at which new information arrives about the present and future observations X and Y, given past observations Z. In an experiment using a very simple Markov-chain model applied to two Philip Glass minimalistic music pieces, Abdallah shows that these measures reveal the structure of the pieces in agreement with the judgment of a human expert listener.

3.7.1.4 Modeling Expectation by Listening to Audio Music Sequences

Dubnov et al. [2007] propose an algorithm to causally build prediction suffix trees so as to describe the redundancy of an audio signal. Hazan et al. [2007] propose a causal and unsupervised system that learns the structure of an audio stream and predicts its continuation. The first step is to discretize a mid-level representation of the signal into a sequence of symbols representing inter-onset intervals, timbre and pitch. Based on the symbolic sequences heard so far, a prediction of the next symbols is made. From this, the system can predict the nature and the timing of the next musical event. The system can be seen as an application of symbolic expectation systems (see 3.7.1.3) to audio and makes it possible to compute similar information-theoretic measures to characterize how the structure of musical excerpts is learned.

4 Conclusion

We have described the processing chain that hierarchically ascends from audio signal transduction to symbolic music cognition. In the beginning of this chain, auditory neuroscience provides a solid ground on the auditory periphery architecture and involved processes. As we progress upwards, more information from other fields is required to get a proper picture: psychoacoustics provides just noticeable differences, thresholds of elementary acoustic features such as pitch and loudness, or masking models; music psychology experiments as well as cognitive musicology investigate grouping principles and musical expectancy. Building explanatory and predictive computational models of human music processing can be faced from different perspectives: the direct neuro-mimetic approach is based on computational models of neurons, and on population dynamics; modeling processes such as categorization and schema activation, unsupervised and recurrent online (statistical) machine learning algorithms provide useful tools; "higher" processes, involving the interaction of symbolic units can be described by cognitive architectures such as ACT-R, or by grammar-based approaches that make evident the relationships between music and language. High-level processes can feedback to low-level processes, e.g. modulate perception, as in the case of attention. Of special interest is the role of motion-related cognitive processes not only during music performance but also during music listening [Leman, 2008]. One obstacle to be overcome by a properly grounded music cognition model is the direct connection of audio music signals to abstract, internal representations (i.e. the "symbolic grounding problem", Harnad [1990]). The "psychological reality" of music scores as a candidate of such abstract representations has to be questioned and challenged. Unsupervised learning, developmental studies, and between-species comparative research are feeding evidences to pursue that goal. Build-

ing music models on a scientifically justified perceptual and cognitive basis will pave the way for understanding and solving applied problems such as those expressed by the following questions: What are the significant similarities between musical excerpts? (music information retrieval) How does musical intelligence and competence develop? (music education) How can we predict the emotional and cognitive reaction of a listener? (emotion engineering) In Part II, we discuss how domain-specific models, e.g. for melody, rhythm, and harmony, can take us further toward the clarification of these questions.

5 Acknowledgment

We would like to thank Joshua Young for useful comments and Graham Coleman for proof reading the manuscript. This work is funded by EU Open FET IST-FP6-013123 (EmCAP) and the Spanish TIC project ProSeMus (TIN2006-14932-C02-01). The first author also received support from a Juan de la Cierva scholarship from the Spanish Ministry of Education and Science. The third author is funded by the Austrian National Science Fund, FWF (project: P19349-N15).

Appendix

A Software for Simulation of Early Auditory Processing

Implementations of auditory periphery and mid-level processing models are contained in the following distributions: The *Auditory Toolbox* (*Matlab*, Slaney [1998a]) implements auditory periphery models as well as sound and speech processing techniques. The *Development System for Auditory Modeling (DSAM)* (C-language, O'Mard et al. [1994, 1997]) enables the user to design an auditory processing chain choosing from various alternative models of the basilar membrane, the inner hair cell, and the auditory nerve. The *IPEM Toolbox* (ipem.ugent.be) features the following components: auditory periphery, roughness, onset extraction, pitch completion, rhythm, echoic memory, and context. *ACT-R/PM* is an extension for the ACT-R cognitive architecture framework (Section 3.7.1, Byrne and Anderson [1998]). ACT-R/PM is the perceptual-motor layer (including a rudimentary auditory module) that quantizes the stimuli and forwards them to the ACT-R cognition layer. It is organized into two separate components: *where* and *what*. The *where* component locates the acoustic stimulus and forwards it to the ACT system. Once ACT decides to focus attention to the acoustic event; the *what* component analyzes it and creates a chunk with specific features (e.g. frequency, timbre class, loudness; Huss and Byrne [2003]).

References

Abbott, A., 2002. Neurobiology: Music, maestro, please! Nature 416 (6876), 12–14.

Abdallah, S., Plumbley, M., 2007. Information dynamics. Tech. Rep. C4DM-TR07-01, Centre for Digital Music, Queen Mary, University of London.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Qin, Y., 2004. An integrated theory of the mind. Psychological Review 111 (4), 1036–1060.

Ayotte, J., Peretz, I., Rousseau, I., Bard, C., Bojanowski, M., 2000. Patterns of music agnosia associated with middle cerebral artery infarcts. Brain 123 (9), 1926–1938.

Barlow, H. B., 1989. Unsupervised learning. Neural Computation 1, 295–311.

Bartlett, F. C., 1932. Remembering: A study in experimental and social psychology. Cambridge University Press, London York.

Besson, M., Friederici, A., 2005. Part II: Language and music A comparison. The neurosciences and music. Annals of the New York Academy of Sciences 1060, 57–58.

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

Bregman, A. S., Dannenbring, G., 1973. The effect of continuity on auditory stream segregation. Perception & Psychophysics 13, 308–312.

Britannica, 2003. Deconstruction. In: Encyclopaedia Britannica.

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

Byrne, M. D., Anderson, J. R., 1998. Perception and action. In: Anderson, J. R., Lebiere, C. (Eds.), The Atomic Components of Thought. Lawrence Erlbaum, Mahwah, NJ, pp. 167–200.

Cardoso, J.-F., 1998. Blind signal separation: statistical principles. In: Proc. of the IEEE, special issue on blind identification and estimation. pp. 2009–2025.

Carlyon, R., Cusack, R., Foxton, J., Robertson, I., 2001. Effects of attention and unilateral neglect on auditory stream segregation. J. of Experimental Psychology: Human Perception and Performance 27, 115–127.

Carnevale, N., Hines, M. L., 2006. The NEURON Book. Cambridge University Press, Cambridge, UK, http://www.neuron.yale.edu/neuron/.

Casey, M. A., Westner, A., 2000. Separation of mixed audio sources by independent subspace analysis. In: Proc. of the Int. Computer Music Conf. ICMA, pp. 154–161.

Chomsky, N., 1988. Language and problems of knowledge. In: Martinich, A. P. (Ed.), The Philosophy of Language. Oxford University Press.

Chowning, J. M., 1980. Computer synthesis of the singing voice. In: Sound Generation in Winds, Strings, Computers. Vol. 29. Royal Swedish Academy of Music, Stockholm, pp. 4–13.

Comon, P., 1994. Independent component analysis, a new concept? Signal Processing 36, 287– 314.

Cooke, M., Brown, G., Ellis, D., Wrigley, S., 1997. The MATLAB auditory and speech demos project. Http://www.dcs.shef.ac.uk/∼martin/MAD/docs/mad.htm.

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

Darwin, C., Ciocca, V., 1992. Grouping in pitch perception: Effects of onset synchrony and ear of presentation of a mistuned component. J. of the Acoustical Society of America 91, 3381–3390.

Darwin, C., Hukin, R., Al-Khatib, B. Y., 1995. Grouping in pitch perception: Evidence for sequential constraints. J. of the Acoustical Society of America 98, 880–885.

De Barreto, G. A., Araujo, A. F. R., Kremer, S. C., 2003. A taxonomy for spatiotemporal connectionist networks revisited: The unsupervised case. Neural Computation 15 (6), 1255–1320.

de Cheveigné, A., 2005. Multiple f0 estimation. In: Wang, D., Brown, G. J. (Eds.), Computational Auditory Scene Analysis. John Wiley.

de la Motte, D., 1980. Harmonielehre, 3rd Edition. Bärenreiter, Basel.

Deliège, I., 1987. Le parallélisme, support d'une analyse auditive de la musique: Vers un modèle des parcours cognitifs de l'information musicale. Analyse musical 6, 73–79.

Deliège, I., 1989. A perceptual approach to contemporary musical forms. Contemporary Music Review 4, 213–230.

Deliège, I., 1992. Paramètres psychologiques et processus de segmentation dans l'écoute de la musique. In: Dalmonte, R., Baroni, M. (Eds.), Secondo convegno europeo di analisi musicale. Universita' degli studi di Trento, pp. 83–90.

Deutsch, D., 1999. The processing of pitch combinations. In: Deutsch, D. (Ed.), The psychology of music, 2nd Edition. Series in Cognition and Perception. Academic Press, pp. 349–411.

Dowling, W. J., 1999. The Psychology of Music, 2nd Edition. Academic Press, Ch. Development of Music Perception and Cognition, pp. 603–25.

Dubnov, S., Assayag, G., Cont, A., 2007. Audio oracle: A new algorithm for fast learning of audio structures. In: International Computer Music Conference. Vol. 2. pp. 224–227, http://books.nips.cc/nips05.html.

Eggermont, J., 2000. Sound-induced synchronization of neural activity between and within three auditory cortical areas. Neurophysiology 83 (5), 2708–2722.

Ehrenfels, C. v., 1890. Über Gestaltqualitäten. Vierteljahrsschrift für wissenschaftliche Philosophie 14, 249–292.

Ellis, D., 1996. Prediction-driven computational auditory scene analysis. Ph.D. thesis, MIT.

Engel, A., Roelfsema, P. R., Fries, P., Brecht, M., Singer, W., 1997. Role of the temporal domain for response selection and perceptual binding. Cerebral

Cortex 7, 571–582.
Farah, M., Rabinowitz, C., 2003. Genetic and environmental influences on the organisation of semantic memory in the brain: Is "living things" an innate category? Cognitive Neuropsychology 20 (3-6), 401–408.
Fisher, D. H., 1987. Knowledge acquisition via incremental conceptual clustering. Mach. Learn. 2 (2), 139–172.
Fishman, Y., Volkov, I., Noh, M., Garell, P., Bakken, H., 2001. Consonance and dissonance of musical chords: neural correlates in auditory cortex of monkeys and humans. J. Neurophysiol. 86 (6), 2761–2788.
Fraisse, P., 1957. Psychology of Time. Presses Universitaires de France, trans. from Psychologie du temps.
Fries, W., Swihart, A., 1990. Disturbance of rhythm sense following right hemisphere damage. Neuropsychologia 28 (12), 1317–1323.
Gjerdingen, R. O., 1988. A Classic Turn of Phrase: Music and the Psychology of Convention. University of Pennsylvania Press, Philadelphia.
Grossberg, S., 1976. Adaptive pattern classification and universal recoding: I. parallel development and coding of neural feature detectors. Biological Cybernetics 23, 121–134.
Hall, D. A., Johnsrude, I., Haggard, M. P., Palmer, A. R., Akeroyd, M. A., Summerfield, A. Q., 2002. Spectral and temporal processing in human auditory cortex. Cereb. Cortex 12, 140–149.
Hargreaves, D., North, A. C., 1997. The Social Psychology of Music. Oxford University Press.
Harnad, S., 1990. The symbol grounding problem. Physica D 42, 335–346.
Hart, H. C., Palmer, A. R., Hall, D. A., 2003. Amplitude and frequency-modulated stimuli activate common regions of human auditory cortex. Cereb. Cortex 13 (7), 773–781.
Haueisen, J., Knoesche, T., 2001. Involuntary motor activity in pianists evoked by music perception. J. of Cognitive Neuroscience 13, 786–792.
Haykin, S., 1999. Neural Networks, 2nd Edition. Prentice-Hall, Upper Saddle River, NJ.
Hazan, A., Brossier, P., Marxer, R., Purwins, H., 2007. What/when causal expectation modelling in monophonic pitched and percussive audio. In: NIPS Music, Brain and Cognition Workshop. Whistler, CA.
Helmholtz, H. v., 1863. Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik. Vieweg, Braunschweig.
Hirsh, I., 1959. Auditory perception of temporal order. J. of the Acoustical Society of America 31, 759–767.
Hirsh, I., Monahan, C., Grant, K., Singh, P., 1990. Studies in auditory timing, 1: Simple patterns. Perception and Psychophysics 47, 215–226.
Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9 (8), 1735–1780.
Hodgkin, A. L., Huxley, A. F., 1952. A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. 117, 500–544.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., Thagard, P. R., 1986. Induction: processes of inference, learning, and discovery. MIT Press, Cambridge, MA, USA.

Huss, D. G., Byrne, M. D., 2003. An ACT-R/PM model of the articulatory loop. In: Detje, F., Doerner, D., Schaub, H. (Eds.), Proceedings of the Fifth International Conference on Cognitive Modeling. Universitas-Verlag, Bamberg, Germany, pp. 135–140.

Ivry, R. B., Keele, S., 1989. Timing functions of the cerebellum. J. Cogn. Neurosci. 1, 136–152.

Janata, P., Grafton, S., 2003. Swinging in the brain: shared neural substrates for behaviors related to sequencing and music. Nat. Neurosci. 6, 682–687.

Justus, T., Bharucha, J., 2001. Modularity in musical processing: the automaticity of harmonic priming. J. Exp. Psychol.: Hum. Percept. Perform. 27, 1000–1011.

Kandel, E., Schwartz, J., Jessell, T., 1991. Principles of Neural Science, 3rd Edition. Prentice-Hall.

Kashino, K., Nakadai, K., Kinoshita, T., Tanaka, H., 1998. Application of the Bayesian probability network to music scene analysis. In: Rosenthal, D. F., Okuno, H. G. (Eds.), Computational Auditory Scene Analysis. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 115–137.

Koelsch, S., Siebel, W. A., 2005. Towards a neural basis of music perception. Trends in Cognitive Sciences 9 (12), 578–584.

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

Krumhansl, C., 2000. Rhythm and pitch in music cognition. Psychol. Bull. 126 (1), 159–179.

Lee, T.-W., Girolami, M., Bell, A. J., Sejnowski, T. J., 2000. A unifying information-theoretic framework for independent component analysis. Computers & Mathematics with Applications 39 (11), 1–21.

Leman, M., 1995. Music and Schema Theory. Vol. 31 of Springer Series in Information Sciences. Springer, Berlin, New York, Tokyo.

Leman, M., 2008. Embodied Music Cognition and Mediation Technology. MIT Press, Cambridge, MA.

Lidov, D., 1999. Elements of Semiotics. Signs and Semaphores. St. Martin's Press.

Liégeois-Chauvel, C., Peretz, I., Babai, M., Laguitton, V., Chauvel, P., 1998. Contribution of different cortical areas in the temporal lobes to music processing. Brain 121, 1853–1867.

Lyon, R., Shamma, S., 1996. Auditory representations of timbre and pitch. In: Auditory computation. Springer, pp. 221–270.

Maass, W., 1997. Networks of spiking neurons: the third generation of neural network models. Neural Networks 10, 1659–1671.

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

Majewska, A. K., Newton, J. R., Sur, M., 2006. Remodeling of synaptic structure in sensory cortical areas in vivo. The J. of Neuroscience 26 (11), 3021–

3029.

Mandler, J. M., 1979. Categorical and schematic organization in memory. In: Puff, C. R. (Ed.), Memory Organisation and Structure. New York: Academic Press, pp. 259–299.

Marxer, R., Holonowicz, P., Purwins, H., Hazan, A., 2007. Dynamical hierarchical self-organization of harmonic, motivic, and pitch categories. In: NIPS Music, Brain and Cognition Workshop. Vancouver, Canada.

McAdams, S., 1989. Psychological constraints on form-bearing dimensions in music. Contemporary Music Review, 181–198.

Meddis, R., Hewitt, M. J., June 1991. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. J. of the Acoustical Soc. of America 89 (6), 2866–2882.

Miller, G. A., 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63, 81–97.

Mozer, M. C., Das, S., 1992. A connectionist symbol manipulator that discovers the structure of context-free languages. In: Neural Information Processing Systems (NIPS). Whistler, Canada, pp. 863–870, http://books.nips.cc/nips05.html.

Müller, K.-R., Philips, P., Ziehe, A., Jan. 1999. JADE-TD: Combining higher-order statistics and temporal information for blind source separation (with noise). In: Proc. of the 1st Int. Workshop on Independent Component Analysis and Signal Separation (ICA-99). Aussios, France, pp. 87–92.

Murata, N., Ikeda, S., Ziehe, A., 2001. An approach to blind source separation based on temporal structure of speech signals. Neurocomputing 41 (1-4), 1–24.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P., Winkler, I., 2001. 'Primitive intelligence' in the auditory cortex. Trends in Neurosciences 24 (5), 283–288.

Nakatani, T., Okuno, H. G., Kawabata, T., 1995. Residue-driven architecture for computational auditory scene analysis. In: Proc. of the 14th Int. Joint Conf. on Artificial Intelligence (IJCAI'95). Vol. 1. pp. 165–172.

Normann, I., 2000. Tonhöhenwahrnehmung: Simulation und Paradoxie. Diploma Thesis, University of Tübingen.

Obermayer, K., Ritter, H., Schulten, K., 1990. A principle for the formation of the spatial structure of cortical feature maps. Proc. of the National Academy of Science USA 87, 8345–8349.

Okuno, H. G., Ikeda, S., Nakatani, T., Aug. 1999. Combining independent component analysis and sound stream segregation. In: Proc. of the IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99). Stockholm, Sweden, pp. 92–98.

O'Mard, L. P., Hewitt, M. J., Meddis, R., 1994. LUTEear: Core Routines Library. Speech & Hearing Laboratory Loughborough University of Technology, Loughborough.

O'Mard, L. P., Hewitt, M. J., Meddis, R., 1997. LUTEar 2.0.9 Manual. http://www.essex.ac.uk/psychology/hearinglab/lutear/manual/Manual.html.

Oppenheim, A. V., Schafer, R. W., 1989. Discrete-Time Signal Processing. Prentice-Hall, Englewood Cliffs, NJ.

Parra, L., Spence, C., 2000. Convolutive blind separation of non-stationary sources. In: IEEE Transactions Speech and Audio Processing. pp. 320–327.

Pearce, M., Wiggins, G., 2004. Improved methods for statistical modelling of monophonic music. J. of New Music Research 33 (4), 367–385.

Penhune, V., Doyon, J., 2002. Dynamic cortical and subcortical networks in learning and delayed recall of timed motor sequences. J. Neurosci. 22 (4), 1397–1406.

Peretz, I., 1990. Processing of local and global musical information by unilateral brain-dam aged patients. Brain 113, 1185–1205.

Peretz, I., Zatorre, R., 2005. Brain organization for music processing. Annual Review of Psychology 56 (1), 89–114.

Peskin, C., 1975. Mathematical Aspects of Heart Physiology. Courant Institute of Mathematical Sciences, New York University.

Pietro, M. D., Laganaro, M., Leeman, B., Schnider, A., 2004. Receptive amusia: temporal auditory deficit in a professional musician following a left temporo-parietal lesion. Neuropsychologia 42, 868–877.

Pinker, S., 1997. How the Mind Works. W. W. Norton, New York.

Platel, H., Baron, J., Desgranges, B., Bernard, F., Eustache, F., 2003. Semantic and episodic memory for music are subserved by distinct neural networks. Neuroimage 20 (1), 244–256.

Platel, H., Price, C., Baron, J., Wise, R., Lambert, J., 1997. The structural components of music perception. Brain 120, 229–243.

Purwins, H., 2005. Profiles of pitch classes - circularity of relative pitch and key: Experiments, models, computational music analysis, and perspectives. Ph.D. thesis, Berlin University of Technology.

Purwins, H., Blankertz, B., Obermayer, K., 2000a. Computing auditory perception. Organised Sound 5 (3), 159–171.

Purwins, H., Blankertz, B., Obermayer, K., 2000b. A new method for tracking modulations in tonal music in audio data format. In: Amari, S.-I., Giles, C. L., Gori, M., Piuri, V. (Eds.), Int. Joint Conf. on Neural Networks (IJCNN-00). Vol. 6. IEEE Computer Society, pp. 270–275.

Purwins, H., Blankertz, B., Obermayer, K., 2008. Toroidal models in tonal theory and pitch-class analysis. In: Hewlett, W. B., Selfridge-Field, E. (Eds.), Computing in Musicology. Vol. 15. Center for Computer Assisted Research in the Humanities and MIT Press, Menlo Park, in print.

Rameau, J. P., 1722. Traité de l'harmonie réduite à ses principes naturels. Ballard, Paris.

Regnault, P., Bigand, E., Besson, M., 2001. Event-related brain potentials show top-down and bottom-up modulations of musical expectations. J. Cogn. Neurosci. 13, 241–255.

Riemann, H., 1877. Musikalische Syntaxis. Breitkopf und Härtel, Leipzig.

Roederer, J. G., 1995. The Physics and Psychophysics of Music, 3rd Edition. Springer.

Rosenthal, D. F., Okuno, H. G. (Eds.), 1998. Computational Auditory Scene Analysis. L. Erlbaum Assoc.

Roskies, A. L., 1999. The binding problem. Neuron 24 (1), 7–9.

Rumelhart, D. E., 1980. Schemata: The building blocks of cognition. In: Spiro, R., Bruce, B., Brewer, W. (Eds.), Theoretical Issues in Reading Comprehension. Lawrence Erlbaum, New Jersey, pp. 33–58.

Sanger, T. D., 1989. Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Networks 2, 459–473.

Scheirer, E. D., 2000. Music-listening systems. Ph.D. thesis, MIT.

Schenker, H., 1935. Der freie Satz. Vol. 3 of Neue musikalische Theorien und Phantasien. Universal, Wien.

Schopenhauer, A., 1859. Die Welt als Wille und Vorstellung. Frankfurt.

Semple, M. N., Scott, B. H., 2003. Cortical mechanisms in hearing. Current Opinion in Neurobiology 2 (13), 167–173.

Shadlen, M., Movshon, J., 1999. Synchrony unbound: a critical evaluation of the temporal binding hypothesis. Neuron 24 (1), 67–77.

Shamma, S., 2001. On the role of space and time in auditory processing. Trends in Cognitive Sciences 5, 340–348.

Slaney, M., 1994. Auditory model inversion for sound separation. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP). Adelaide, Australia, pp. 563–569.

Slaney, M., 1998a. Auditory toolbox. Tech. Rep. 1998-010, Interval Research Corporation, http://rvl4.ecn.purdue.edu/∼malcolm/interval/1998-010/.

Slaney, M., 1998b. Connecting correlograms to neurophysiology and psychoacoustics. In: Palmer, A. R., Rees, A., Summerfield, A. Q., Meddis, R. (Eds.), Psychophysical and Physiological Advances in Hearing. Whurr Publishers.

Snyder, B., 2000. Music and Memory: An Introduction. The MIT Press, Cambridge, MA.

Terhardt, E., 1974. Pitch, consonance and harmony. J. of the Acoustical Soc. of America 55, 1061–1069.

Terhardt, E., Stoll, G., Seewann, M., 1982. Algorithm for extraction of pitch and pitch salience from complex tonal signals. J. of the Acoustical Soc. of America 71 (3), 679–688.

Terman, D., Wang, D. L., 1995. Global competition and local cooperation in a network of neural oscillators. Physica D 81, 148–176.

Tillmann, B., Bharucha, J., Bigand, E., 2001. Implicit learning of regularities in Western tonal music by self-organization. In: Proceedings of the Sixth Neural Computation and Psychology Conference. Springer, pp. 175–184.

Todd, P. M., 1991. A connectionist approach to algorithmic composition. In: Todd, P., Loy, D. (Eds.), Music and Connectionism. MIT Press, pp. 173–194.

Tramo, M. J., 2001. Music of the hemispheres. Science 291 (5501), 54–56.

Treisman, A., 1999. Solutions to the binding problem: Progress through controversy and convergence. Neuron 24 (1), 105–125.

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

Vignolo, L. A., 2003. Music agnosia and auditory agnosia. Ann. NY Acad. Sci. 999, 50–57.

Wang, D. L., 1996. Primitive auditory segregation based on oscillatory correlation. Cognitive Science 20, 409–456.

Warren, J., Uppenkamp, S., Patterson, R., Griffiths, T., 2003. Separating pitch chroma and pitch height in the human brain. Proc. Natl. Acad. Sci. USA 100 (17), 10038–10042.

Wilson, S., Pressing, J., Wales, R., 2002. Modelling rhythmic function in a musician post-stroke. Neuropsychologia 40, 1494–1505.

Wrigley, S., Brown, G., 2005. A computational model of auditory selective attention. IEEE Transactions on Neural Networks 15 (5).

Yost, W. A., Hill, R., 1979. Pitch and pitch strength of ripple noise. J. of the Acoustical Soc. of America 66, 400–410.

Zatorre, R., 1985. Discrimination and recognition of tonal melodies after unilateral cerebral excisions. Neuropsychologia 23, 31–41.

Zatorre, R., 1988. Pitch perception of complex tones and human temporal-lobe function. J. of the Acoustical Society of America 84:2, 566–572.

Zwicker, E., Fastl, H., 1999. Psychoacoustics – Facts and Models, 2nd Edition. Springer.

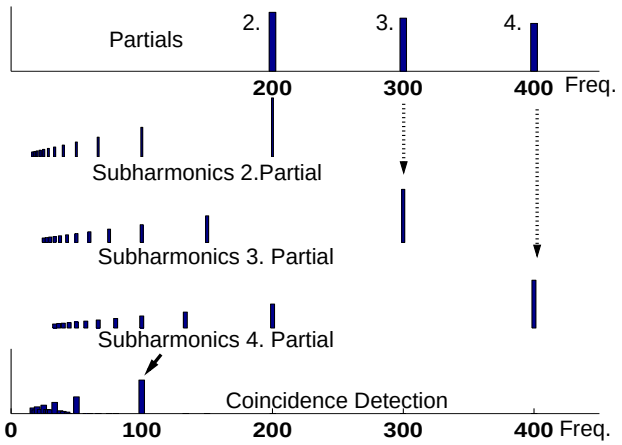

Fig. A.1. Detection of a missing fundamental by the recognition of regularities in the spectral peak pattern induced by the partials. Consider a tone (top row) with missing fundamental frequency $f_0 = 100$ Hz composed of partials with frequencies $2f_0, 3f_0, 4f_0$. For each partial f_i, a subharmonic series with frequencies $\frac{f_i}{m}$ and decreasing energy is constructed. In the further step of coincidence detection (bottom line), the accumulation of the subharmonic series leads to the spectrum of possible pitches. Its maximum (arrow) presents the (most prominent) pitch, in this case 100 Hz.

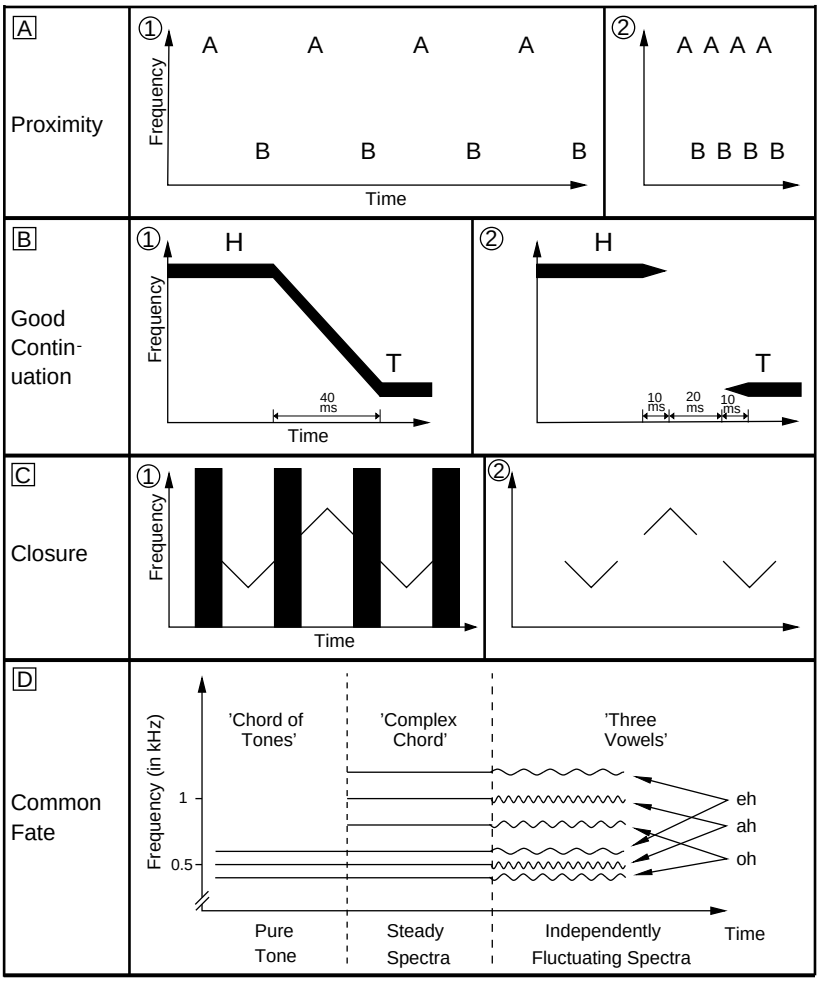

Fig. A.2. Psychoacoustic experiments demonstrating grouping principles, Figure according to Bregman [1990]).

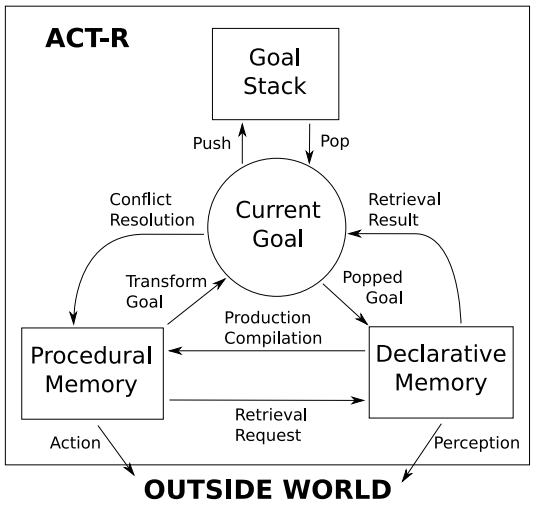

Fig. A.3. Flow of information among the various modules of ACT-R (Adapted from Anderson & Lebiere, 1998).