Hendrik Purwins [1], Benjamin Blankertz [2] , Klaus Obermayer[1]

Modelle der Musikwahrnehmung zwischen auditorischer Neurophysiologie und Psychoakustik

10. Juni 2000

Einleitung

Der herkömmliche Weg der Musikanalyse fußt auf dem Notentext. Einzelne Noten bilden einen Akkord. Akkordfolgen definieren ein tonales Zentrum. Aus den tonalen Zentren des Stückes ergibt sich seine Tonart. Die Repräsentation von Musik als Notentext entspricht nur zum Teil den Entitäten der neuronalen Verarbeitung auditorischer Wahrnehmung im Gehirn. Wie können wir diese internen Repräsentationen von Musik finden? Zwei Wege steuern auf dieses Ziel von unterschiedlichen Seiten zu. Auf der einen Seite ist die auditorische Neurophysiologie, die die Funktionsweise von Haarzellen und auch von späteren Verarbeitungsstufen untersucht. Den zweiten Weg bilden die Psychoakustik und die Musikpsychologie. Insbesondere die Gestaltprinzipien, die isolierte auditorische Wahrnehmungen einer akustischen Quelle zuordnen, helfen hier weiter. Die Konvergenz beider Gebiete vollzieht sich langsam. Einige Fortschritte werden durch Übertragung von Erkenntnissen über das visuelle System gewonnen, gegenüber denen die auditorische Forschung zurücksteht. Es entstehen neue wissenschaftliche Disziplinen, die die vorliegenden Erkenntnisse mathematisch beschreiben sowie Computersimulationen dazu durchführen. In der Anwendung sind biologisch oder psychologisch inspirierte Algorithmen bisher meist pragmatischen Heuristiken unterlegen.

Zum Aufbau des Artikels: Nach der Beschreibung der Reizleitung in Neuronen, kommen wir zum Aufbau des Ohres und der Hörbahn. Es folgen eine Diskussion der auditorischen Gestalt-Prinzipien, der Neuronen- und auditorischen Modelle, dann der Implementierung der Gestalt-Prinzipien insbesondere mit ICA sowie der Anwendungsbeispiele: Mit einer logarithmisch auflösenden Filterbank können wir Modulationsverläufe in einem Musikstück verfolgen (Purwins et al. [2000]). Als Beispiel für ein kognitives Modell werden wir zeigen, daß in einem auditorischen Modell in Kombination mit der selbstorganisierenden Merkmalskarte (Kohonen [1982]) zwischentonartliche Relationen gut repräsentiert werden (Leman

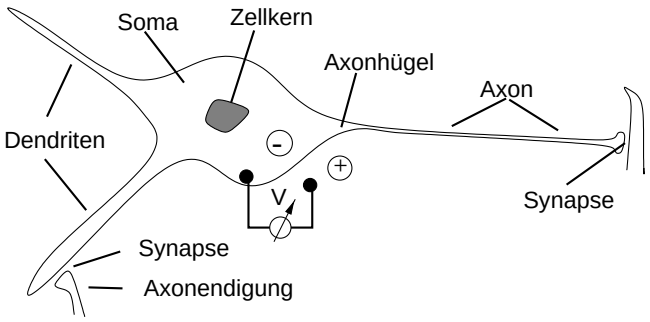
[1]Sekr. FR 2-1, FB 13, Technische Universit"at Berlin, Franklinstr. 28/29, 10 587 Berlin, {hendrik,oby}@cs.tu-berlin.de

[2]GMD FIRST, Rudower Chaussee 5, 12489 Berlin, blanker@first.gmd.de

[1995], Blankertz et al. [1999]). Invertierte auditorische Modelle können zur Kodierung benutzt werden (Slaney [1994]).

1 Neurophysiologie

1.1 Spikende Neuronen



In der Mitte des Neurons befindet sich im Soma der Zellkern. Vom Soma ausgehende Verzweigungen, mit denen das Neuron Information empfangen kann, heißen Dendriten. An einer Stelle, dem Axonhügel, buchtet sich das Neuron ein wenig aus. Von dort geht das Axon aus. Die Stelle, an der sich die sendende Zelle mit einer Axonenendigung an einen Dendriten der empfangenden Zelle anlagert, heißt Synapse.

Reizweiterleitung Das Ruhepotential wird für eine Zelle im unerregten Zustand zwischen Innerem und Äußerem der Zelle gemessen und liegt bei -70 mV. Wie ist nun der Signalfluß im Neuron? Die Synapse injiziert Strom in den Dendriten. Der dadurch erzeugte Spannungspuls wird durch Ionendiffusion an der Membran des Dendriten weitergeleitet. Der Puls klingt exponentiell ab, bis er den Axonhügel erreicht. Dort werden die Spannungspulse bis zum Erreichen eines Schwellenwertes akkumuliert. Dies löst im Axon einen stereotypen impulsartigen Spannungsverlauf (*Spike*) aus. Im Axon geschieht die Reizleitung an der Membran entlang durch Öffnen und Schließen von Ionenkanälen für Na^+ und K^+.

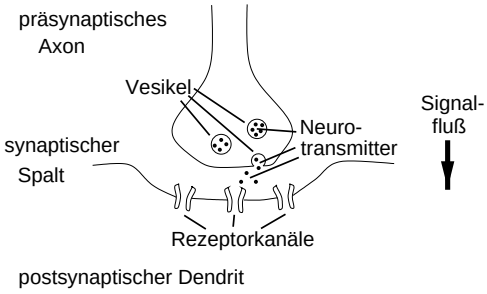


ABBILDUNG 2: Die Synapse

Im präsynaptischen Axon der sendenden Zelle sammeln sich Neurotransmitter in Vesikeln (Abb. 2). Die Vesikel schütten die Neurotransmitter in den synaptischen Spalt, abhängig von den eintreffenden Spannungspulsen. Die emittierten Neurotransmitter werden über Rezeptorkanäle vom der postsynaptischen Dendriten der empfangenden Zelle aufgenommen.

Excitatorische (*inhibitorische*) Synapsen reagieren auf einen Spannungspuls in der präsynaptischen Endigung mit einem Spannungsanstieg, d.h. einer *Depolarisation* (Spannungsabfall, d.h. *Hyperpolarisation*).

Zeitkodierung von Neuronen Wir konnten sehen, daß bei einer Depolarisation über einen Schwellenwert über das Ruhepotential hinaus, ein Spike zu beobachten ist. In einer Folge von Spikes kann dann wie folgt Information kodiert werden: (i) durch die exakten Zeitpunkte, wann gefeuert wird (*Zeitkodierungshypothese*), (ii) durch die zeitlichen Abstände zwischen zwei aufeinanderfolgenden Spikes (Interspike-Intervall) oder (iii) durch die *Spikerate* (Kehrwert des mittleren Interspike-Intervalls).

Merkmalsbindung durch synchrones Spiken Das *Bindungsproblem* besteht darin, wie eine Reihe von Merkmalen ein akustisches oder visuelles Objekt konstituieren. Engel et al. [1993] gibt ein Modell zur Lösung des Bindungsproblems im Visuellen an: Die Zeitkodierungshypothese nimmt an, daß Objekte im visuellen Cortex jeweils von synchron spikenden Neuronen repräsentiert werden. Ein visuelles Objekt wird dann durch ein Assembly von Neuronen dargestellt, die elementare Objektmerkmale detektieren. Die Zusammengehörigkeit der Merkmale wird durch die zeitliche Korrelation der Neuronenaktivitäten eines Assemblies abgebildet. Diejenigen Neuronen, die zum selben Assembly gehören, feuern nach der Zeitkodierungshypothese jeweils synchron.

1.2 Das Ohr

Akustische Reize werden durch das Ohr aufgenommen, in elektrische Signale überführt und anschließend entlang der Hörbahn weiterverarbeitet.

Außen- und Mittelohr (Abb. 3)

Die Schallwellen werden durch eine Verstärkerkaskade, bestehend aus äußerem Gehörgang und Mittelohr (mechanische Verstärkung durch die Resonanzeigenschaften des Außenohres sowie *Hammer*, *Amboß* und *Steigbügel*), auf das *ovale Fenster* der Schnecke (Cochlea) des Innenohrs übertragen. Die Innenohrschnecke wird durch die *Reissner'sche* und die *Basilarmembran* längs in drei Räume aufgeteilt. Eine Auslenkung des ovalen Fensters führt zu einer Wanderwelle, die im untersten Raum bis ans Ende hinaufläuft und im obersten Raum zurückkehrt.

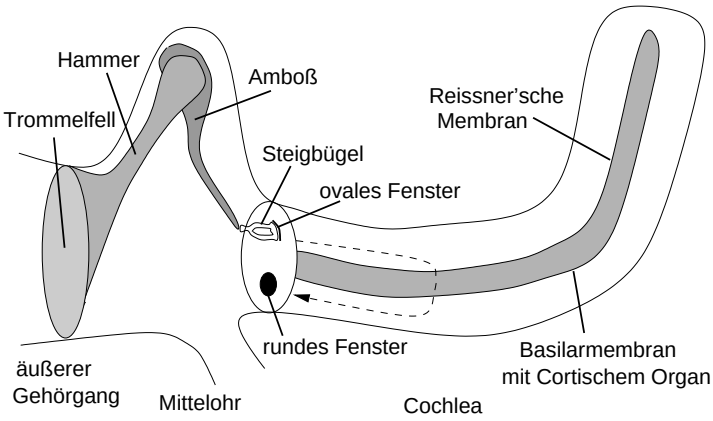
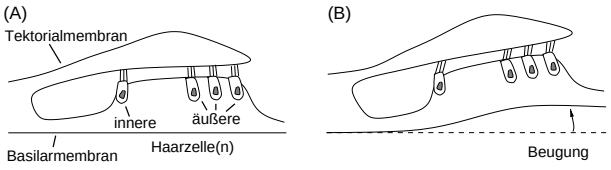


ABBILDUNG 3: Die Cochlea

Haarzelle Auf der Basilarmembran sitzt das Cortische Organ, das aus ca. 3000 Haarzellen besteht.



Diese Haarzellen werden durch die Wellenbewegung der Basilarmembran ausgelenkt(B). Die Haare der Haarzellen werden gebeugt, was zu einer von der Amplitude der Auslenkung abhängigen elektrischen Aktivität führt. Die Resonanzfrequenz der Haarzelle ist bedingt durch die Steifigkeit der Basilarmembran an der Stelle und durch die Steifigkeit, Größe, und elektrische Resonanz der Haarzelle selbst. Frequenzen werden logarithmisch auf die Basilarmembran abgebildet (Weber-Fechnersches Gesetz). Das im auditorische Bereich herrschende Prinzip der *Tonotopie* besagt, daß benachbart liegende Haarzellen auf benachbarte Neuronen im Zentralnervensystem projezieren.

Bei einer Frequenz des Signals von bis zu 4-5 kHz werden die Haarzellen nur in der Phase erregt, in der sie „aufwärts" gebogen werden (Abb. 1.2). Unterhalb dieser Frequenzen treten Spikes in der Haarzelle fast nur während der positiven Phase des Signals auf (*Phaselocking*).

1.3 Die Hörbahn

Die Signale der Haarzellen werden durch den *Hörnerv* an den *Hörkern* weitergeleitet, unter Berücksichtigung des „Tonotopie-Prinzips".

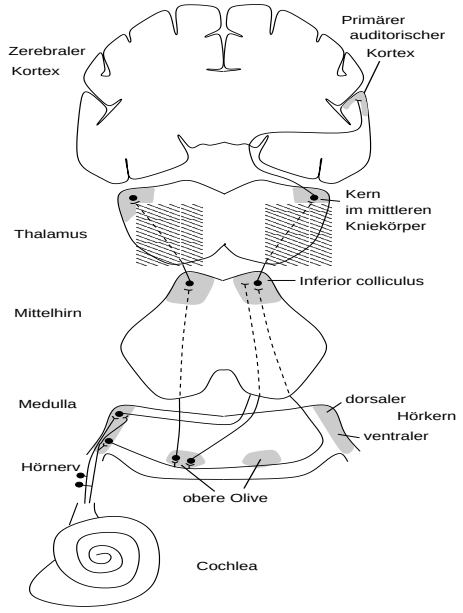


ABBILDUNG 5: Die Hörbahn

Im Hörkern findet man verschiedene Typen von Zellen: Bushy-, Stellate- und Fusiform-Zellen (Oertel [1983]). Bushy-Zellen indizieren den Beginn, Stellate-Zellen das Andauern und Amplitudeninformation und Fusiform-Zellen das Ende eines akustischen Signals.

Durch efferente Nervenfasern, die unter anderem vom oberen Olivenkomplex kommen, werden die Dämpfung der Empfindlichkeit des auditiven Systems, und vielleicht auch Mechanismen zum Heraushören wichtiger Signale aus dem Umgebungslärm gesteuert (Prinzip „Figur und Hintergrund", s.u.). Durch neuronale Rückverbindungen können die Haarzellen in Bewegungen versetzt werden, die daraufhin ihrerseits wieder über die Cochlea und das Mittelohr Klang erzeugen (*otoakustische Emissionen*).

Von den Hörkernen führt die Hörbahn zum *oberen Olivenkomplex*, in dem es Nervenzellen gibt, die zur Lokalisation von Schallquellen im Raum beitragen können. Zur Lokalisation tragen einerseits Nervenzellen für niederfrequente Töne bei, die auf interaurale Laufzeitunterschiede ansprechen und andererseits Nervenzellen für hochfrequente Töne, die auf interaurale Intensitätsunterschiede reagieren.

Vom oberen Olivenkomplex aus verlaufen Nerven zur Formatio reticularis, zu den *Kernen im mittleren Kniekörper* und zum *Inferior Colliculus*. Im Colliculus inferior und auch an anderen Stellen der Hörbahn findet man Nervenzellen, die auf amplitudenmodulierte akustische Signale ansprechen. Die Hörbahn verläuft weiterhin zum *primären auditorischen Kortex* und von dort aus in höhere kortikale Areale . Auch in diesen Arealen finden sich tonotope Karten.

2 Psychoakustik

2.1 Tonhöhenwahrnehmung

Tonhöhe kann definiert werden als „die Eigenschaft der auditorischen Wahrnehmung, auf Grund derer Klänge in eine musikalische Skala eingeordnet werden können." (American Standards Association, 1960) Im folgenden soll von Tonhöhe immer als von einem Wahrnehmungsattribut die Rede sein.

Im einfachen Fall besteht ein harmonischer Ton aus Grundfrequenz und Obertönen, deren Frequenzen Vielfache der Grundfrequenz sind. Ausgehend von einem solchen Ton entfernte Terhardt et al. [1982] sukzessive die Grundfrequenz und die tiefen Obertöne. Die wahrgenommene Tonhöhe blieb gleich. Die Klangfarbe hingegen wurde heller. Eine solche *„virtuelle Tonhöhe "* wird nur dann gehört, wenn mindestens ein Oberton im Frequenzbereich 500-1500 Hz liegt und benachbarte Obertöne nicht zu eng beieinander liegen (Terhardt [1972], Terhardt [1992]). Bei einem harmonischen Ton ist die virtuelle Tonhöhe der größte gemeinsame Teiler seiner Frequenzkomponenten. Zur Berechnung der virtuellen Tonhöhe kann die Autokorrelationsfunktion (s.u.) verwendet werden.

Wenn die geraden Obertöne etwas schwächer sind als die ungeraden, kann es sein, daß eine Grundfrequenz gehört wird, die doppelt so hoch ist, wie der Abstand zwischen zwei benachbarten Obertönen. Bei den meisten Instrumenten weichen die Obertöne gering von den ganzzahligen Vielfachen einer Grundfrequenz ab. Ist die Frequenzverteilung der Obertöne jedoch zu unregelmäßig (*unharmonisches Spektrum*), wie z.B. bei Glockenklängen, so ist die Tonhöhenwahrnehmung bei unterschiedlichen Personen sehr uneinheitlich.

2.2 Maskierung

Es gibt Maskierung im Zeitbereich und im Frequenzbereich.

Frequenzbereich Man hat ein schmalbandiges Rauschen (Maske) einer gegebenen Lautstärke. Dann wird ein Sinustön mit benachbarter Frequenz gespielt. Der Sinuston ist jetzt unhörbar. Der Frequenzabstand wird so lange vergrößert, bis der Sinuston gerade hörbar ist. Dies wird mehrere Male für unterschiedlich laute Sinustöne wiederholt. So erhält man die Maskierungskurve des schmalbandigen Rauschens, die angibt, welche Sinustöne - abhängig von ihrer Frequenz und dem Frequenzabstand zum Rauschen - von der Maske „verschluckt " werden. Über dieses Verfahren bekommt man die kritischen Bandbreiten (s.u.). Ist die Maske ein Sinuston, erhält man ähnliche Kurven.

Zeitbereich Ein Klang maskiert auch vorangehende und nachfolgende Töne. Die Maskierungskurve erstreckt sich bis auf 20ms vor und 200ms nach dem Signal und läßt sich durch Exponentialfunktionen beschreiben.

2.3 Gestalt-Prinzipien

Der Mensch ist ständig einem großen Wirrwar von Sinneseindrücken ausgesetzt. Wie können wir Objekte in unserer Umwelt ausmachen? Das *Bindungsproblem* besteht darin, daß einzelne sensorische Informationen nach bestimmten Kriterien eine *Gestalt* bilden, die etwas über das Objekt der Außenwelt aussagt.

Als Ursprungstext der Gestaltpsychologie kann das 1890 vom Freiherrn Christian von Ehrenfels verfaßte Werk „Über Gestaltqualitäten" (von Ehrenfels [1890]) gelten. Anregung sowie die Verwendung des Begriffs „Gestalt" rührt von Werken des Physikers Mach [1886] her. Es sollte noch einige Jahre dauern, bis systematische Forschung zur Formulierung der Gestaltgesetze führte (siehe z.B. Wertheimer [1923]). Obwohl sich schon Mach und von Ehrenfels bei ihren Erörterungen auch explizit auf Musikwahrnehmung bezogen hatten, widmeten sich die nachfolgenden Studien besonders der Untersuchung der visuellen Wahrnehmung. Das Aufkommen computergestützter Klangsynthese und -analyse wurde verstärkt seit den siebziger Jahren für psychoakustische Experimente genutzt, um die Gestaltgesetze in Bezug auf die akustische Wahrnehmung zu präzisieren und zu ergänzen (*Auditory Scene Analysis, (ASA)*).

Im folgenden sollen sieben wichtige akustische Gestalt-Prinzipien kurz vorgestellt werden (Moore [1989] S. 244-253, Bregman [1990] S. 18-29,S. 196-202): (A I) Nähe, (A II) Ähnlichkeit, (B) Gute Fortsetzung, (C) Geschlossenheit, (D) Gemeinsames Schicksal, (E) Disjunkte Zuordnung von Klangquellen, (F) Figur und Hintergrund sowie Aufmerksamkeit.

In der visuellen Wahrnehmung gruppiert das Prinzip der „*Nähe*" (A I) Elemente, die räumlich nah benachbart liegen. Dabei ist „nah " relativ gemeint. Elemente, die zu einer Gruppe zusammengefaßt werden, müssen untereinander einen kleineren Abstand haben, als zu Elementen einer anderen Gruppe. Das Prinzip der Nähe kann auch auf die Wahrnehmung von Tönen übertragen werden, wenn man es auf die Abstände ihrer Einsatzzeiten, Tonhöhen oder Lautstärken bezieht.

Folgen Töne gleicher Tonhöhe schnell aufeinander, getrennt durch Pausen von vorangegangenen und folgenden Tönen, werden sie zusammen gruppiert. Wenn zwei aufeinander folgende Töne (relativ zu anderen) einen kleinen Tonhöhenabstand untereinander haben werden sie meist gruppiert. Zeitliche und Tonhöhennähe sind konkurrierende Kriterien. So wird die langsame Tonfolge A-B-A-B... (Abb. 2.3 A 1)), die große Intervallsprünge enthält, als eine Line wahrgenommen. Bei starker Beschleunigung und gleichbleibender Tonhöhe hört man die Linie der A's und eine zweite Linie der B's separat.

Das Gestalt-Prinzip der *„Ähnlichkeit"* (A II) ist dem Prinzip der „Nähe" (A I) sehr ähnlich. Bregman [1990] S. 198 bezieht es auf Eigenschaften eines Klanges, die sich nicht so einfach auf eine physikalische Dimension zurückführen lassen. Hören wir ein Musikinstrument eine Melodie spielen, so gruppieren sich die Töne aufgrund der Ähnlichkeit der Klangfarbe zu der Melodie.

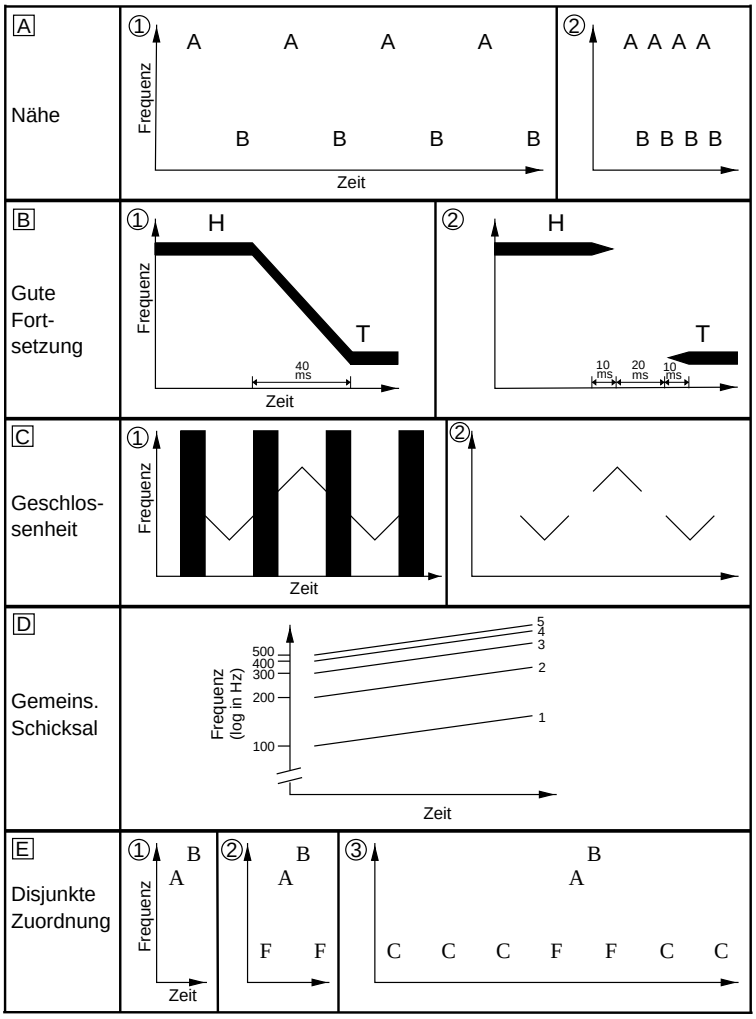


ABBILDUNG 6: Gestalt-Experimente, vgl. Text

Es ist die physikalische Eigenschaft einer Klangquelle, daß sich im allgemeinen Änderungen der Frequenz, der Lautstärke, der Richtung und des Spektrums stetig und allmählich vollziehen und nicht abrupt. Das Prinzip *„Gute Fortsetzung "* (B) ordnet einer kontinuierlichen Veränderung solcher Merkmale eine sich verändernde Quelle zu. Plötzliche Veränderungen hingegen zeigen das Auftreten einer neuen Quelle an. In Bregman and Dannenbring [1973] (Abb. 2.3 B) werden abwechselnd hohe (H) und tiefe (T) Töne gespielt. Werden alle Töne mit Glissandi verbunden (Abb. 2.3 B 1), gruppieren sich alle Töne zu einer Linie. Bleiben hohe und tiefe Töne unverbunden (Abb. 2.3 B 2), so gruppieren sich H's und T's jeweils zu einer eigenen Linie. Das Prinzip „Gute Fortsetzung" ist der stetige Spezialfall

des Prinzips „Nähe" für beliebig kleine Änderungen.

Für die Gestaltpsychologen gibt in der visuellen Wahrnehmung bestimmte „gute Gestalten", z.B. Kreise. Sind diese nur fragmentarisch zu sehen, „schließt" die Wahrnehmung die Gestalt ab. Dieses „*Prinzip der Geschlossenheit*" (C) kommt auch in folgendem Experiment (Abb. 2.3 C) zum tragen: Ein auf- und absteigendes Glissando wurde durch Pausen unterbrochen (Abb. 2.3 C 2). Man hört drei abgesetzte Linien. Dann wurden die Pausen mit Rauschen gefüllt (Abb. 2.3 C 1), das so laut war, daß es das Glissando maskiert hätte, w'are es nicht unterbrochen. Es wird ein durchgehendes Glissando wahrgenommen. Das unterbrochene Glissando hat eine „gute Gestalt", da die Glissandi vor und nach der Pause auf Grund der Tonhöhennähe zu einer Gestalt gruppiert und durch eine „wahrgenommene gute Fortsetzung " ergänzt werden, dies jedoch nur dann, wenn ein Störgeräusch detektiert wurde, das eine hypothetische Fortsetzung maskieren würde. Diese Ergänzung kann auch als auditorischer Kompensationsmechanismus für die Maskierung interpretiert werden.

Das Prinzip „*Gemeinsames Schicksal*" (D) tritt in Erscheinung, wenn bei einem aus mehreren Frequenzkomponenten zusammengesetzten Klang ähnliche Veränderungen zur gleichen Zeit vor sich gehen. Wenn Frequenzkomponenten synchron vibrieren oder glissandieren (vgl. Abb. 2.3 D), so werden sie gruppiert zu einem Ton mit einer besser heraushörbaren Tonhöhe und Klangfarbe. Auch synchrone Amplitudenmodulation gruppiert Frequenzkomponenten zu einem Ton ebenso wie synchrone On- und Offsets. Wenn zwei Instrumente mit einer Zeitverzögerung von 30ms einsetzen, können sie schon als unterschiedliche Stimmen aufgefaßt werden. Dies erleichtert das Verfolgen von polyphoner Musik

Das Prinzip der „*Disjunkten Zuordnung von Klangquellen*" (E) besagt, daß Klangmerkmale entweder dem einen oder dem anderen musikalischen Objekt zugeordnet werden und selten zwei musikalischen Objekten zugleich angehören können. In Abb. 2.3 E 1 wird A-B als ein Motiv gruppiert. In Abb. 2.3 E 2 werden die tiefen Flankentöne F auch in die Gruppe mit einbezogen. Schließlich gehen die Flankentöne F in Abb. 2.3 E 3 wegen des gleichförmigen Rhythmus und der gleichen Frequenz mit den C's in einer Gruppe auf, während A B zusammen eine separate Gruppe bilden.

„*Figur und Hintergrund*" bezeichnet das Phänomen, daß bei einer Cocktail-Party die Aufmerksamkeit auf einen einzelnen Sprecher gelenkt werden kann, während alles andere als Hintergrund erscheint.

Vorwissen bzw. Kontext spielt bei der Objekterkennung ebenfalls eine wichtige Rolle.

2.4 Musikpsychologie

Tonarten entsprechen Histogrammen von Chromata (Abb. 7) Die „*Probe Tone*" Experimente wurden von Krumhansl and Shepard [1979] entwickelt. „*Probe Tone Profile*" sind eine quantitave Beschreibung einer Tonart. Dieses Konzept er-

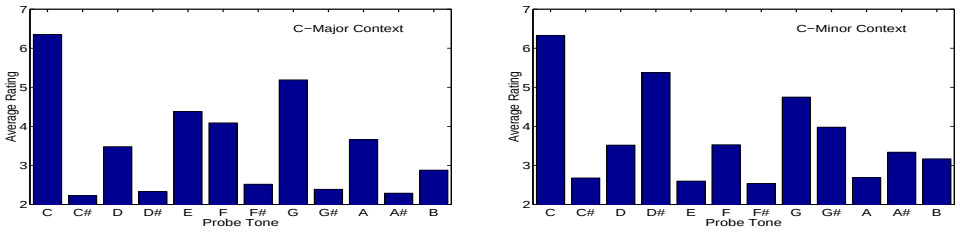


ABBILDUNG 7: Probe Tone Profile geben eine Hierarchie von Chromata für jede Tonart vor. Ein tonaler Kontext von C-Dur (c-moll) wird etabliert, durch Vorspielen einer Kadenz. Dann wird die Versuchsperson gefragt, wie gut unterschiedliche Chromata zu dem tonalen Kontext passen. Die Abbildung zeigt die durchschnittlichen Antworten von Versuchspersonen mit durchschnittlich 7,5 Jahren Musikerziehung (Krumhansl and Kessler [1982],das englische B ist das deutsche H).

laubt es, statistische und Computeranalysen von Musik einerseits und Kognitionspsychologie andererseits wechselseitig aufeinander zu beziehen (Blankertz et al. [1999]).

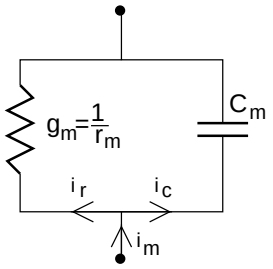
In einer tonalen Komposition einer gegebenen Tonart korrespondiert jede Komponente im entsprechenden Probe Tone Profil eng mit der Gesamthäufigkeit und Gesamtdauer der Chromata auf prominenten Taktzeiten. Die Bedeutung einer Tonart ergibt sich aus ihrem Bezug zu allen anderen Tonarten. Diese Bezüge lassen sich quantifizieren durch Berechnung der Korrelation oder des Euklidischen Abstands zwischen den Profilen.

3 Mathematische Beschreibung

3.1 Computational Neuroscience

Reizleitung im Dendriten Um den Stromfluß durch einen Dendriten zu beschreiben, kann man den Dendriten durch ein zylindrisches Kabel approximieren, das durch Durchmesser und Länge bestimmt ist.

Nach dem 1. Kirchhoffschen Gesetz kann man aus dem Ersatzschaltbild für das Kabel den „*Leaky Integrator*" bekommen:



$$C_m \frac{\partial}{\partial t} V(t) = -g_m * V(t) + i_m,$$

wobei i_m ein externes Signal, z.B. ein synaptischer Strom ist. g_m beschreibt das Leck durch die Membranleitfähigkeit. Komplexe Dendritenstrukturen können stückweise durch Zylinder angenähert werden.

Reizleitung im Axon Es gilt nach dem 1.Kirchhoffschen Gesetz für den Gesamtstrom I_m:

$$I_m(t) = C_m \frac{\partial V(t)}{\partial t} + I_{Na} + I_K + I_{leak}.$$

Darin bezeichnet C_m die Kapazität der Membran, die durch ihre Permeabilitätseigenschaften bestimmt wird. I_K ist der Strom, der durch den Fluß von K^+-Ionen induziert wird, I_{Na} der für Na^+-Ionen, I_{leak} ist der „Verlust ", d.h. der Strom, der durch die Membran an den Ionenkanälen vorbei fließt.

Hieraus ergibt sich ein System von teils nicht-linearen Differentialgleichungen, die die zeitabhängige Leitfähigkeit der Na^+- und der K^-Kanäle beschreiben. Dies ist das *Hodgkin-Huxley-Modell* (Hodgkin and Huxley [1952]), der Ausgangspunkt der meisten heute entwickelten Neuronen-Modelle zur aktiven Reizleitung.

Zeitkodierung Auf dem Kodierungsprinzip der Spikerate beruht das *konnektionistische* Neuron, und somit die Mehrheit der klassischen Künstlichen Neuronalen Netze. Zur komplexeren Modellierung des Spannungsverlaufs und des Spikeverhaltens kann man das System von Differentialgleichungen von Hodgkin-Huxley, unter einer vereinfachenden Geometrie mit Nebenbedingungen lösen.

Um diese Aufgabe zu vereinfachen, gibt es jedoch noch andere Modelle. Im „Integrate and Fire-" Modell wird die Stromstärke bis zum Erreichen des Schwellenwertes aufintegriert. Dann wird der stereotype Verlauf der Spikephase einfach eingesetzt. Nach der Refraktärzeit wird dann wieder vom Ruhepotential aus begonnen aufzuintegrieren.

3.2 Modelle der Hörbahn

Die Vorverabeitung akustischer Reize durch das Ohr und die Hörbahn ist entscheidend für die zentral stattfindende Analyse und für die anschließende Wahrnehmung der Signale. Zum einen werden die Daten für die höheren Verarbeitungsleistungen aufbereitet, zum anderen werden relevante Signale bzw. Signalparameter extrahiert. Die biologienahe Modellierung der frühen akustischen Verarbeitung im Computermodell hilft dem Verständnis und läßt sich zur Analyse und Datenkompression nutzen.

Bandpaß für Außen- und Mittelohr Der Effekt von Außen- und Mittelohr wird durch einen oben abgeflachten Bandpaß (Nedzelnitsky [1980]) als rückgekoppelter (IIR-) Filter zweiter Ordnung implementiert. Das Ergebnis ist eine Antwortkurve, die für tiefe und hohe Frequenzen abfällt und ihr Maximum bei 2000 Hz hat.

Basilarmembran als Filterbank Die Basilarmembran wird von einer Filterbank implementiert. Das eindimensionale Schalldrucksignal wird auf viele Kanäle aufgeteilt, die jeweils nur Signalanteile eines sehr beschränkten Frequenzbereichs enthalten. Die Ausgabe wird als *Cochleogramm* bezeichnet. Zunächst bietet

sich die Diskrete Fouriertransformation an. Diese löst im Frequenzbereich überall gleichmäßig auf. Der Wahrnehmung näher kommt eine Filterverteilung, die sich gleichmäßig über den logarithmierten Frequenzbereich verteilt. Für nicht allzu tiefe Frequenzen ist hier die die Constant Q Transformation (Brown [1991], Brown and Puckette [1992]) ein geeignetes Verfahren. Auch die stetige Wavelet Transformation (CWT) kommt in Frage. Berücksichtigt man Maskierungseffekte im Frequenzbereich, so kommt man zu „Kritischen Bandbreiten" (CBU) oder „Equivalenten Rechteckigen Bandbreiten" („equivalent rectangular bandwidth", ERB). Die Verteilungsdichte der Filter beschreiben die CBUs durch eine Funktion, die bis 500 Hz linear, dann logarithmisch ist. Noch genauer mit experimentellen Daten stimmt die ERB-Rate Funktion überein. Sie ist eine in x- und y-Richtung verschobene logarithmische Funktion. Die Filterform wird gut durch Gammatone-Filter (Moore and Glasberg [1983]) angenähert.

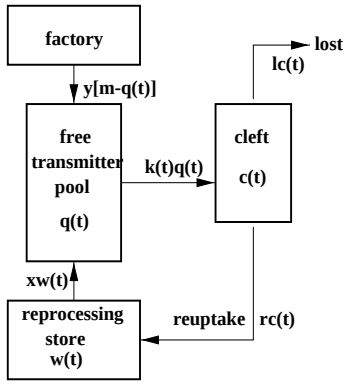
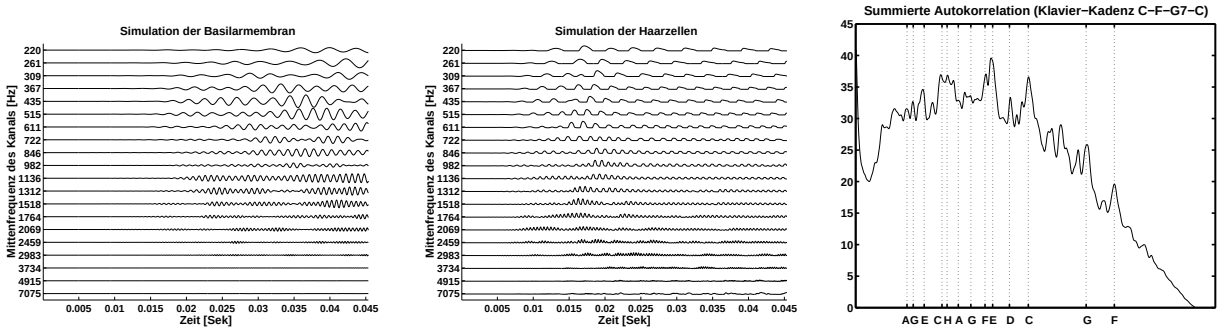


ABBILDUNG 9: Haarzellenmodell, vgl. Text

Es kann als weitere Verarbeitung ein Modell der Haarzellensynapse folgen. Exemplarisch wird hier das Modell von Meddis and Hewitt [1991] erläutert. In dem Modell (vgl. Abb. 3.2 (a)) wird die Erregung, die diese Zelle an den Hörnerven weitergibt, als proportional zur Anzahl der Transmitter im synaptischen Spalt $c(t)$ angenommen. $c(t)$ hängt von der Anzahl der Transmitter $q(t)$ in der Haarzelle über die (nichtlineare) Funktion $k(t)$ ab. $k(t)$ beschreibt die Durchlässigkeit der präsynaptischen Haarzellenwand an der Synapse und wird vom Membranpotential getriggert. Je näher $q(t)$ an der Maximalkapazität m liegt, desto weniger Transmitter werden neu produziert. Außerdem läuft ein Teil (Faktor r) der Transmitter aus dem synaptischen Spalt wieder zurück und benötigt einige Zeit (Faktor x), um wieder zur Verfügung zu stehen. Die einzelnen Bestandteile des Modells sind die „Fabrik" der Transmitterstoffe, der Transmitterpool, der Synaptische Spalt, und der Rücklaufspeicher. Das zeitliche Verhalten des Systems wird durch entspechende nichtlineare Differentialgleichungen erster Ordnung beschrieben: Die Änderung eines Parameters in jeder Box errechnet sich aus der Bilanz der ein- und ausfließenden Größen, z.B. $\frac{\partial c}{\partial t} = k(t)q(t) - lc(t) - r\,c(t)$.

Dieses Haarzellenmodell reproduziert einige wichtige Eigenschaften von Haarzellen von Meerschweinchen (die den menschlichen Haarzellen ähnlich sind): (i) Übertragungsfunktion von Intensität nach Spikerate (ii) Adaptionsverhalten bei kurzen Signalen: die Haarzelle spikt beim Einsetzen eines Tones stark. Ihre Feuerungsrate nähert sich beim Aushalten des Tones einem konstanten Wert und fällt nach dem Ende des Tons wieder ganz ab (Westerman [1985]). (iii) Phaselocking: nimmt mit steigender Frequenz zwischen 1 kHz und 5 kHz stark ab .

Die zeitliche Folge der Vektoren der Feuerungsraten $s(t)$ sind die Ausgabe des Haarzellenmodells.



(a) Basilarmembran (b) Haarzelle (c) Autokorrelationsfunktion

ABBILDUNG 10: Eine C-Dur Klavier-Kadenz durchläuft das auditorische Modell

Autokorrelation zur Grundtonbestimmung Zu jedem Zeitpunkt t berechnet man für die durch die Fensterfunktion $w_t(k)$ modifizierte Autokorrelationsfunktion

$$R_i(n,t) = ((w_t \cdot s_i)(k) * s_i(-k))(n)$$

$$= (\mathcal{F}^{-1}(\mathcal{F}(w_t \cdot s_i)(k) \cdot \mathcal{F}s_i(-k)))(n) = \sum_k w_t(k)s_i(k)s_i(k+n),$$

wobei $\mathcal{F}$ die Fouriertransformation, $w_t(k)$ eine Rechteckfunktion oder eine Rechteckfunktion multipliziert mit $\exp(\frac{k}{T})$ ist mit der Abklingkonstanten T. $R_i(n,t)$ gibt für jeden Kanal i an, wie stark die Grundperioden und ihre Vielfachen der einzelnen Frequenzkomponenten sind, aus denen sich das Signal $s_i(t)$ zusammensetzt. Berechnen läßt sich $R_i(n,t)$ gut durch Verwendung der Fast Fourier Transformation, einen sehr effizienten Algorithmus zur Berechnung der Fouriertransformation.

Die Autokorrelationen $R_i(n,t)$ der einzelnen Komponenten werden schließlich über alle Kanäle i summiert zu $R_n(t)$. Für unterschiedliche n entsprechen die $R_i(n,t)$ unterschiedlichen Periodenlängen und ihren Vielfachen. Im Gegensatz zur einfachen Fouriertransformation wird die Frequenzauflösung bei der Autokorrelation für tiefe Frequenzen immer besser. Die Interpretation eines solchen Vektors ist jedoch nicht so klar wie bei der einfachen Fouriertransformation (vgl. Abb. 10 c)).

Da in $R_n(t)$ die Grundperioden und ihre Vielfachen von $R_i(n,t)$ aufaddiert werden, entspricht das Maximum in $R_n(t)$ dem kleinsten gemeinsamen Vielfachen der Grundperioden der einzelnen Frequenzkomponenten. In Frequenzen gesprochen, ist dies der größte gemeinsame Teiler, somit die virtuelle Tonhöhe des Klanges. Diese Technik ist eine Verallgemeinerung des „Subharmonic Sum Spectrum" in

Terhardt [1992]. Leman [1994] sieht die Autokorrelation auch biologisch motiviert, indem er auf die Experimente von Schreiner and Langner [1988] verweist, die möglicherweise tonotope Organisation auch für Neuronen erkennen lassen, die auf amplitudenmodulierede Signale reagieren. Scheirer hält Repräsentation durch *Korrelogramme* (Bilder die durch die Autokorrelation erzeugt wurden) für eine mögliche wahrnehmungsrelevante musikalische Entität im Gegensatz zur Notenschrift. Er reproduziert im Experiment auch einige Phänomene der Auditory Scene Analysis. Meddis dagegen hält die Autokorrelationsfunktion für biologisch unplausibel [3].

Tonotope kortikale Karten Die Selbstorganisierende Merkmalskarte (*SOM*) wurde als Modell neuronaler Lernprozesse von Kohonen [1982] vorgeschlagen und hat seitdem eine weite Verbreitung als biologienahes Modell und als Algorithmus für die unüberwachte Merkmalsextraktion gefunden.

Eine SOM besteht aus (in der Regel) einer Schicht konnektionistischer Neuronen, die effektiv über kurzreichweitig erregende und langreichweitig hemmende Verbindungen miteinander wechselwirken. Jedes Neuron ist mit jeder Komponente des Eingabevektors verbunden, wobei jeder Verbindung ein synaptisches Gewicht w zugeordnet ist. Wird ein Eingabemuster x präsentiert, so berechnet sich die Ausgabeaktivität y_j des j-ten Neurons der Neuronenschicht zu

$$(1) \quad y_j = \begin{cases} 1 \text{ falls } j = \text{argmax}_i |x - w_i| \\ 0 \text{ sonst.} \end{cases}$$

Der Algorithmus: (i) Zufällige Initialisierung der Gewichte w_i, (ii) zufällige Auswahl eines Datenvektors x, (iii) Winner-takes-all Schritt: $i(x) = arg_j min \|x - w_j\|$, (iv) Lernschritt:

$$w_j = \begin{cases} w_j + \nu[x - w_j] : j \in \Lambda_{i(x)} \\ w_j \qquad\qquad : \text{sonst} \end{cases},$$

wobei ν die Lernrate und $\Lambda_{i(x)}$ die Nachbarschaftsfunktion ist, mit dem Gewinnerneuron $i(x)$ als Zentrum. ν, und $\Lambda_{i(x)}$ werden während des Trainings graduell verkleinert. (v) Weiter bei (ii).

Nach dem Training haben sich die Gewichte jedes Neurons auf einen anderen Teil des Datenraumes „spezialisiert“, so daß in der Neuronenschicht benachbarte Neurone in Bezug auf die euklidische Metrik auf ähnliche Eingabedaten reagieren. Der Algorithmus setzt das Prinzip der Tonotopie um.

Die Topographie-Eigenschaft, die ja wesentlich für die SOM ist, kann auch mit informationstheoretischen Methoden erreicht werden durch das Prinzips des „Maximalen Erhalts von Information” (Linsker [1989]).

[3] Persönliches Gespräch

Zeitliche Kontextverarbeitung Gedächtnis kann mit dem „Leaky Integrator" modelliert werden. Eine solche Kontextverabeitung entspricht der Betätigung des Hallpedals beim Klavierspiel.

Auf einer größeren Zeitskala können vorverarbeitete Musiksignale mit Hidden Markov Modellen verarbeitet werden.

3.3 Computational Auditory Scene Analysis (CASA)

Das Ziel der CASA ist es, die in der ASA gewonnenen Erkenntnisse in Algorithmen umzusetzten und als Computerprogramme zu implementieren.

Oszillatormodelle als Modell zur Lösung des Bindungsproblems Die Lösung des Bindungsproblems durch zeitliche Kodierung kann durch Oszillatorsysteme implementiert werden (Wang [1998]).

Onset-Detektion Für ein Zellmodell, in dem ein Onset detektiert wird , kann das Membranpotential $p(t)$ wie folgt modelliert werden (Brown and Cooke [1998]):

$$p(t) = p(t-1)c + E_{psp}\, r(t) - I_{psp}\, r(t-1)$$

E_{psp} und I_{psp} sind die Stärken der exitatorischen und inhibitorischen Inputs. $r(t)$ ist das Signal der Haarzelle und c bestimmt, wie schnell $p(t)$ wieder zurück auf das Ruhepotential fällt. Die Feuerungsrate für diese Zelle ist $s(t) = max(p(t), 0)$.

„Figur und Hintergrund" Eine wichtige Aufgabe dieses Bereichs stellt die akustische Quellentrennung dar. Dabei soll ein Gemisch von mehreren Klangquellen (bestehend aus Sprache, Musik und anderen Geräuschen), das durch ein oder mehrere Mikrophone aufgenommen wird, in seine ursprünglichen Bestandteile zerlegt werden. Da eine wichtige Anwendung in der Horgerättechnik liegt, ist es wünschenswert mit höchstens zwei Mikrophonen auszukommen.

Obwohl es viele erfolgversprechende Ansätze gibt, die Quellentrennung für künstliche Mischungen (Quellen werden einzeln aufgenommen und durch gewichtete Addition digital gemischt) erfolgreich bewältigen, stellt der reale Fall immer noch ein ernsthaftes Problem dar. Die verschiedenen Herangehensweisen können grob in zwei Kategorien eingeteilt werden: (1) starke Orientierung am menschlichen Vorbild und (2) Verwendung von Techniken aus der digitale Signalverarbeitung ohne Blick auf die Biologie. Bestrebungen (1) und (2) synergetisch zu kombinieren werden in Hiroshi G. Okuno [1999] verfolgt. Ein mögliche Vorgehensweise für (1) ist wie folgt. Zur Vorverarbeitung der akustischen Signale wird ein auditorisches Modell benutzt. Dann werden in der Ausgabe des Ohrmodells (Korrelogramme) harmonische Teilstrukturen aufgespürt und daraus spektrale Einheiten gemäß der Gestaltprinzipien gebildet. Diese Einheiten kann man entweder direkt in Klang zurückverwandelt durch übliche Syntheseverfahren (Nakatani et al. [1995]),

oder man bestimmt die spektralen Einheiten als Korrelogrammfolgen und läßt sie die Transformationen des Ohrmodells rückwärts durchlaufen (Slaney [1994],s.u.).

Die größte Gruppe von Verfahren unter (2) beschäftigt sich mit Realisierungen der *Independent Component Analysis* (ICA), (Comon [1994],Cardoso [1998],Müller et al. [1999]). Obwohl Sanger [1989] auch für Ansätze dieser Art einen Zusammenhang zu biologischen Systemen herstellt, sind sie doch eher rein statistische Modellierungen. Idealisiert wird das Problem wie folgt modelliert. Die gesuchten Signale $s_1(t), \ldots, s_n(t)$ werden gemäß einer zeitlich konstanten Mischung linear in m Sensorsignale $x_1(t), \ldots, x_m(t)$ transformiert, die den Ausgabesignalen der Mikrophone entsprechen. (Dabei ist t der Zeitindex, der im folgenden weggelassen wird.) Mit $s = (s_1, \ldots, s_n)^T$ und $x = (x_1, \ldots, x_m)^T$ läßt sich dieser Vorgang in Matrixschreibweise durch $x = As$ darstellen, wobei A die $n \times m$ Mischungsmatrix ist. Die Aufgabe besteht darin, eine Entmischungsmatrix W zu bestimmen, so daß für $y = Wx$ die Komponenten von y den Quellsignalen s entsprechen[4]. Der Ansatz der ICA dies Problem zu lösen basiert auf der Annahme, daß die Quellsignale s_i statistisch unabhängig verteilt[5] sind. Wenn es mindestens soviele Mikrophone wie Quellen gibt ($n \leq m$) und die Mischungsmatrix A invertierbar ist, läßt sich eine Entmischungsmatrix bestimmen. Während die Annahmen der Unabhängigkeit und Invertierbarkeit für die Anwendung auf reale Daten unkritisch sind, stellt es ein Problem dar, daß die Modellierung der Mischung keine realen Echoeffekte berücksichtigt, und daß eine hohe Anzahl von Mikrophonen gebraucht wird. Es reichen nicht zwei Mikrophone für „Figur" und „Hintergrund". Für jede einzelne Klangquelle des Hintergrunds wird ein weiteres Mikrophon benötigt.

Der Bereich der *Audiory Scene Analysis* (ASA) beschäftigt sich mit der Erforschung des Bindungsproblems bezüglich der akustischen Wahrnehmung: Wie zerlegt unsere Wahrnehmung ein akustisches Gemisch in sinnvolle auditorische Einheiten, und wodurch sind diese Einheiten charakterisiert? Welche Mechanismen lassen uns gezielt einen akustischen Fokus setzen?

4 Anwendungen

Die Filtercharakteristik des Außen- und Mittelohres sowie eine nachgeschaltete Filterbank finden Eingang in zahlreiche Aufbauten zur Sprach- und Musikanalyse sowie in gängige Kompressionsverfahren zur Übertragung von Musik (z.B. MPEG-3). In speziellen biologisch / psychoakustisch inspirierten Verfahren werden noch Neuronenmodelle (insbesondere der Haarzelle) mit unterschiedlichen Spike-Repräsentationen, korrelationsbasierte Verfahren, Lernalgorithmen (unüberwachtes Lernen, Hidden Markov Modelle) oder kleine Regelsysteme (black board systems) eingesetzt.

[4] Prinzipiell können Reihefolge und Skalierungen von s nicht bestimmt werden.

[5] und bis auf höchstens eins nicht Gaussverteilt

Constant Q Filterbank als einfaches auditorisches Modell Die constant Q Filterbank kann dazu benutzt werden, um sogenannte *Constant Q Profile* zu berechenen (Brown and Puckette [1992], Izmirli and Bilgen [1996], Purwins et al. [2000]). Ein 12-dimensionales Constant Q Profil steht in enger Beziehung zu Chromata und Probe Tone Profilen. Sie können effizient berechnet werden, sie sind stabil in Bezug auf die Aufnahmequalität der verwendeten Musik und sie sind transponierbar. Constant Q Profil Analyse wird auch als kognitives Modell verwendet. Das Verfahren verwendet als einziges musikalisches Vorwissen Oktaväquivalenz und wohltemperierte Stimmung - minimale Information über tonale Musik. Die relevanten zwischentonartlichen Verwandtschaften entstehen allein auf der Grundlage der trainierten Musikbeispiele (WTK 1 & 2, Chopin Prelúdes op. 28) in einer realen Aufnahme.

Ein Tonarten-Modell aus auditorischer Vorverarbeitung und SOM In dieser Simulation (Leman [1995]) werden einfache Kadenzen der Form I-IV-V7-I in allen Dur- und moll-Tonarten benutzt.

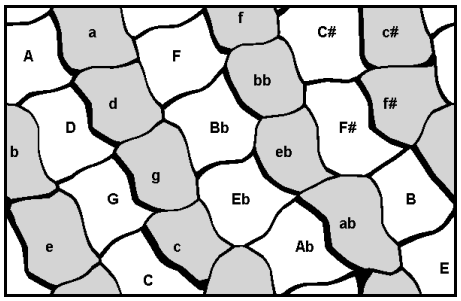


ABBILDUNG 11: Die SOM läßt den Quintenzirkel, die Parallel- und die Obermediantverwandtschaft erkennen (vgl. Text).

Es werden nur *Shepard-Töne* benutzt. Shepard-Töne enthalten Partialtöne, die in Oktavabständen angeordnet sind. Auf der Frequenzachse beschreibt die Hüllkurve der Amplituden eine Kurve, die für eine Frequenz in der Mitte maximal und für hohe und tiefe Frequenzen jeweils abflacht, z.B. gaußförmig oder von abgeflachter Dreiecksform. Da man Shepard-Tönen, zwar ein Chroma zuordnen kann, jedoch keine absolute Tonhöhe, muß das Trainingsmaterial nicht alle Kombinationen von Akkordumkehrungen enthalten. In unserem Aufbau (vgl. Abb. 10 (a)-(c)) werden die digital vorliegenden Kadenzen vorverarbeitet durch folgende Stationen: (1) Digitale (IIR-) Filter, die Außen- und Mittelohr modellieren, (2) logarithmische Filterbank aus 20 Filtern für die Basilarmembran (Abb. 10 (a)), (3) Synapsenmodell der Haarzelle (Abb. 10 (b)), (4) Autokorrelation zur Bestimmung des virtuellen Grundtons (Abb. 10 (c)), (5) Leaky-Integrator als Gedächtnis, (6) die SOM zur Modellierung der Tonotopie.

Nach dem Training durch die SOM organisieren sich die Korrelgramme so, daß wichtige tonartliche Verwandtschaften sichtbar werden (vgl. 11). Große Buchstaben bezeichen Dur, kleine Buchstaben moll. Nach Vorverarbeitung durch ein auditorisches Modell, Autokorrelation und Leaky Integrator werden 24 Dur- und moll-Shepard-Kadenzen als Trainingsmenge benutzt. Die hier benutzte SOM besteht aus einem torusförmigen Netz (21×12 Neurone) (Den Torus erhält man, indem man entgegengesetzten Seiten aneinanderklebt). H und B erscheinen in der englischen Schreibweise als B bzw Bb. Wir können einen Vergleich zwischen den

Probe Tone Profilen (Abb. 7) und dem Autokorrelogram (Abb. 10 (c)) herstellen.

Soundkompression durch Inversion auditorischer Modelle Wir betrachten ein Ohrmodell mit folgendem Aufbau: (1) Filterbank der Cochlea, (2) Abschneiden der negativen Phase durch Gleichrichten, (3) Autokorrelation. Die Ausgaben der dritten Stufe (Korrelogramme) werden vielfach als gut zu interpretierende Darstellung von akustischen Signalen angesehen. Sie eröffnen z.B. einen Ansatz zur Quellentrennung, durch den das Gestaltprinzip ›Figur und Hintergrund‹ technisch umgesetzt wird.

In der Kodierung gilt: Was nicht gehört wird, braucht nicht übertragen zu werden. Solange des Verhältnis Signal/Rauschen (SNR), das durch die Kodierung entsteht, größer ist, als das Verhältnis Signal/Maskierung, ist diese Kodierung akzeptabel. Durch Doppelblindversuche, in denen das Original und die aus dem komprimierten Signal dekodierte Version verglichen werden, kann die Qualität der Codierungsmethode bestimmt werden.

Auf der ERB-Skala haben Maskierungskurven im Frequenzbereich approximativ eine Dreiecksform. Die Maskierungskurve eines komplexen Frequenzgemisches läßt sich nicht einfach linear aus den Masken der Einzelfrequenzen kombinieren. Präzisere Modellierung der Maskierungskurven verwenden auditorische Modelle (Baumgarte [1997]). Maskierungskurven helfen dabei, die zur Verfügung stehenden Bits effektiv zur Kodierung einzelner Frequenzkomponenten aufzuteilen.

Um über Maskierung hinaus weitere auditorische Mechanismen für die Kodierung nutzen zu können, kann man einen Algorithmus konstruieren, der die Verarbeitungsschritte des auditorischen Modells umkehrt (Slaney [1994]). Methoden zur Inversion der Filterbank (1) sind aus der Filtertheorie bekannt. Die Filterung der negativen Phase (2) stellt zunächst ein Problem dar, weil bei diesem Verarbeitungsschritt Information verloren geht. Hier hilft das Verfahren der konvexen Projektionen aus der Bildrekonstruktion, welches es ermöglich apriori Wissen über das ursprüngliche Signal zu nutzen, um verlorene Information möglichst gut zu rekonstruieren. Da jedes Eingabesignal für Schritt (2) (durch die vorgeschaltete Filterbank) stark frequenzbeschränkt ist, liegt geeignetes Wissen über das zu rekonstruierende Signal vor, um diese Methode erfolgreich anzuwenden. Als letztes wird eine Umkehrung der Autokorrelation benötigt. Auch dies scheint zunächst kritisch zu sein, da in der Autokorrelation gegenüber der Fouriertransformation die Phaseninformation nicht enthalten ist. Trotzdem läßt sich das Ausgangssignal im wesentlichen zurückgewinnen. Für den allgemeinen Fall wird in S. H. Nawab [1983] ein rekursives Verfahren angegeben. Wenn es – wie in unserem Fall – um frequenzbeschränkte Signale geht, kann man auch hier mit konvexen Projektionen arbeiten und dadurch bessere Resultate erzielen (Daniel W. Griffin [1984], Slaney [1994]).

Musikanalyse unter Gestaltgesichtspunkten Es werden kompositorische Besonderheiten in der Musikliteratur mit Gestaltgesetzen in Zusammenhang gebracht. So spielt das Prinzip der Nähe eine wichtige Rolle beim Heraushören verschiedener Stimmen im Kontrapunkt.

Webern unterstreicht durch seine Instrumentation des Ricercar aus Bachs „Das Musikalische Opfer" musikalische Gesten, indem mit einer neuen Phrase auch die Instrumentation wechselt. So kann das Prinzip der Ähnlichkeit nicht helfen, die Phrasen als eine Melodie zusammenzufassen. Jede Phrase ist für sich herausgehoben. Dennoch ist die Melodie auch in den vierstimmigen Teilen noch heraushörbar.

Diskussion

Das auditorische System ist noch nicht sehr gut verstanden. Die meisten auditorischen Modelle implementieren nur die allerersten Stadien der auditorischen Verarbeitung grob. Die SOM stellt als Kortex-Modell eine extreme Vereinfachung dar, die lediglich das Tonotopie-Prinzip implementiert. Mehr Wissen über die auditorische Verarbeitung von Klängen wird nötig sein, um eine fundierte Hypothese zur Repräsentation von Musik im Kortex zu formulieren.

Es tun sich einige Probleme im Zusammenhang mit psychoakustischen und musikpsychologischen Experimenten auf. Es ist nicht klar, wie sich die unter künstlichen experimentellen Bedinungen (Sinustöne, oftmalige Wiederholungen der Experimente) erzielten Erkenntnisse auf die komplexere Wirklichkeit übertragen lassen.

Ein weiteres Problem ist die Frage, ob eine Versuchsperson eine starke musikalische Vorbildung hat, oder „naiv " ist. Im Probe Tone Experiment z.B. vermuten wir, daß die Versuchspersonen mit starkem musikalischen Hintergrund, im Probe Tone Experiment explizit ihr Wissen über Skalentöne anwenden. Trotzdem scheint die Probe Tone Methode zur Untersuchung von tonalen Strukturen in nicht-westlicher Musik geeignet zu sein (indische in M. A. Castellano [1984] und koreanische in [Nam 1998]).

Der Erstautor wurde von der „Studienstiftung des deutschen Volkes" and „Axel Springer Stiftung" unterstützt. Wir danken Hauke Bartsch, Mark Leman, Immanuel Normann, Craig Sapp, Cornelius Weber und Gregor Wenning für ihre Hilfe .

Literatur

Baumgarte, F. (1997). A physiological ear model for auditory masking applicable to perceptual coding. In *103rd AES Convention*, New York.

Blankertz, B., Purwins, H., and Obermayer, K. (1999). Toroidal models of interkey relations in tonal music. In *VI. International Conference on Systematic and Comparative Musicology*.

Bregman, A. S. (1990). *Auditory Scene Analysis*. MIT Press, Cambridge, Massachusetts.

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

Brown, G. J. and Cooke, M. (1998). *Computational Auditory Scene Analysis*, chapter Temporal Synchronization in a Neural Oscillator Model of Primitive Auditory Stream Segregation, pages 71–85. L. Erlbaum Assoc.

Brown, J. (1991). Calculation of a constant Q spectral transform. *J. Acoust. Soc. Am.*, 89(1):425–434.

Brown, J. C. and Puckette, M. S. (1992). An efficient algorithm for the calculation of a constant Q transform. *J. Acoust. Soc. Am.*, 92(5):2698–2701.

Cardoso, J.-F. (1998). Blind signal separation: statistical principles. In *Proceedings of the IEEE, special issue on blind identification and estimation*.

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

Daniel W. Griffin, J. S. L. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243.

Engel, A. K., König, P., and Singer, W. (1993). Bildung repräsentationaler zustände im gehirn. *Spektrum der Wissenschaften*, 9:42–47.

Hiroshi G. Okuno, Shiro Ikeda, T. N. (1999). Combining independent component analysis and sound stream segregation. In *Proc. of IJCAI-99 Workshop on Computational Auditory Scene Analysis (CASA'99)*, pages 92–98, Stockholm, Sweden.

Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, 117:500–544. London.

Izmirli, O. and Bilgen, S. (1996). A model for tonal context time course calculation from acoustical input. *Journal of New Music Research*, 25(3):276–288.

Kohonen, T. (1982). Analysis of a simple self-organizing process. *Biol. Cybern.*, 44:135–140.

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.

Krumhansl, C. L. and Shepard, R. N. (1979). Quantification of the hierarchy of tonal function with a diatonic context. *Journal of experimental psychology: Human Perception and Performance*.

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

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

Linsker, R. (1989). How to generate ordered maps by maximazing the mutual information between input and output signals. *Neural Computation*, 1:402–411.

M. A. Castellano, J. J. Bharucha, C. L. K. (1984). Tonal hierarchies in the music of north India. *Journal of Experimental Psychology*, 113(3):394–412.

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

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):2866–2882.

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

Moore, B. and Glasberg, B. (1983). Suggested formulae for calculating auditory filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74:750–753.

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

Nakatani, T., Okuno, H. G., and Kawabata, T. (1995). Residue-driven architecture for computational auditory scene analysis. In *Proc. of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, volume 1, pages 165–172.

Nam, U. (1998). Pitch distributions in Korean court music: Evidence consistent with tonal hierarchies. *Music Perception*, 16(2):243–247.

Nedzelnitsky (1980). Sound pressures in the basal turn of the cat cochlea. *The Journal of the Acoustical Society of America*, 68:1676–1689.

Oertel, D. (1983). Synaptic responses and electrical properties of cells in brain slices of mouse anteroventral cochlear nucleus. *Journal Neuroscience*, 3:2040–2053.

Purwins, H., Blankertz, B., and Obermayer, K. (2000). A new method for tracking modulations in tonal music in audio data format. In *International Joint Conference on Neural Networks*. accepted.

S. H. Nawab, T. F. Quatieri, J. S. L. (1983). Signal reconstruction from short-time fourier transform magnitude. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31:986–998.

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

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

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

Terhardt, E. (1972). Zur Tonhöhenwahrnehmung von Klängen. I. Psychoakustische Grundlagen. *Acustica*, 26:173–186.

Terhardt, E. (1992). Zur Tonhöhenwahrnehmung von Klängen. II. Ein Funktionsschema. *Acustica*, 26:187–199.

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

von Ehrenfels, C. (1890). Über Gestaltqualitäten. *Vierteljahresschrift Wiss. Philos.*, 14:249–292.

Wang, D. (1998). *Computational Auditory Scene Analysis*, chapter Stream Segregation Based on Oscillatory Correlation, pages 71–85. L. Erlbaum Assoc.

Wertheimer, M. (1923). Untersuchungen zur Lehre der Gestalt II. *Psychologische Forschung*, 4:301–350.

Westerman, L. A. (1985). Adaptation and recovery of auditory nerve responses. Special Report ISR-S-24, Institute for Sensory Research, Syracuse University, Syracuse, NY.