Grauel, Becker and Belli (Eds.) Proc. 4th Internat. Workshop "Fuzzy-Neuro-Systems'97, FNS'97, Paderborn/Soest, pp. 442-49, infix-Verlag 1997

Analog-digitaler Chip für ein burstpropagierendes Neuronales Netz

Richard Izak, Karsten Trott, Thomas Zahn, Uwe Markl
Technische Universität Ilmenau
Postfach 100565
D-98684 Ilmenau
Tel.: (03677) 691170
Fax.: (03677) 691163
E-Mail: richard.izak@inf-technik.tu-ilmenau.de

Abstract: This paper describes an approach of biological motivated neural structure for separation of desired information from miscellaneous information sources. The functionality of this structure will be proved in an application for acoustical analysis based on binaural hearing. The aspired demonstrator should adapted itself to an unknown acoustical environment to separate and classify the existing acoustical sources. The neural components of this demonstrator are implemented in a VLSI-CMOS process. Possible fields of application are guidance of intelligent video control systems and acoustical focus in hearing aid.

Stichworte: akustische Signalverarbeitung, analog-digitaler Chip, burstpropagierendes neuronales Netz, VLSI Implementierung

1 Verarbeitung der akustischen Signale

Das Demonstratorsystem kann als eine aus den in Abb. 1 dargestellten Funktionseinheiten bestehende Struktur angesehen werden. Das vom Mikrofon aufgenommene akustische Signalgemisch wird in der Peripheral Processing-Einheit mit Hilfe einer neural gesteuerten Filterbank für eine weitere Analyse im Frequenzbereich vorbereitet. Das gesamte Spektrum des Eingangsignals wird dabei in mehrere überlappende Frequenzbänder unterteilt. Durch eine phasen- und amplitudenrichtige binäre Kodierung mittels Spike-Impulsfolgen (Burst) in 16 Kanälen pro Ohr wird der gesamte Informationsgehalt an das Directional System weitergeleitet. Die Anzahl der Einzelspikes pro Burst ist proportional der Amplitude, während die Phasenlage durch den Zeitpunkt des jeweils ersten Spikes pro Periode definiert wird.
Für die Zugehörigkeit einzelner Frequenzanteile zu einer akustischen Quelle wird die zeitliche Änderung des Frequenzspektrums betrachtet. Die Korrelation zwischen Frequenzverschiebungen der verschiedenen Amplitudenmaxima im Spektrum läßt darauf zurückschliessen, welche Grund- und Oberwellen zu einem Quellensignal gehören. Wesentlich ist, daß in diese Dynamik neben der Amplitude der einzelnen Spektralanteile auch deren Phasenlage in Relationen zueinander bestimmend eingeht.

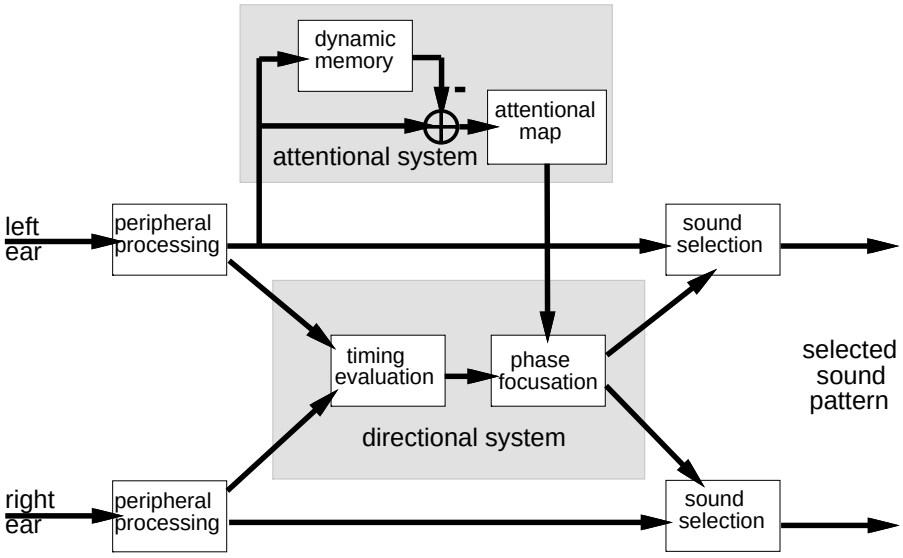

Abbildung 1: Akustisches Aufmerksamkeitssystem

In der Timing-Evaluation Einheit wird durch eine Kreuzkorrelation zwischen Links und Rechts in jedem der Frequenzkanäle die relative Phasenverschiebung und damit die Entstehungsrichtung der Signalanteile ermittelt. Durch Überlagerung der Korrelationsfunktionen aller Frequenzbänder ergeben sich ein oder mehrere prominente Phasenverschiebungen, die der Anzahl und Detektionsrichtung der Signalanteile entsprechen. Somit kann dieses Zwischenergebnis als eine Lokalisierung aller vorhandenen Quellen angesehen werden. Das Aufmerksamkeitsystem dient zur Selektion einzelner Geräusche. Dazu wird das detektierte Signalspektrum mit den im dynamischen Speicher gelernten und bekannten Mustern akustischer Signale verglichen. Nach der Subtraktion bekannter Erregungsmuster am Speicherausgang wird das verbleibende Spike-Muster mittels Koinzidenzdetektoren in eine topologische Struktur überführt. Diese erlaubt deren Identifikation in mehreren symbolischen Ebenen. Eine dauerhafte (10 bis 100μs) währende Aktivierung dieser Karte führt zur Aktivität des Aufmerksamkeitsknotens, welcher zur Phase-Focusation Einheit zurückprojeziert und dort das Feuern der Phasendetektoren ermöglicht. Beide Verarbeitungspfade werden in der Phase-Focusation Einheit verknüpft. Diese ermöglicht mit Hilfe einer lateralen Inhibition das Durchsetzen einer bevorzugten Hörrichtung und leitet zeitgenaue Impulse an die Selektionseinheit weiter. Dort werden mit Hilfe gezielter Disinhibition diejenigen Frequenzanteile aus dem Gesamtmuster extrahiert, welche die bevorzugte Phasenverschiebung aufweisen, also aus der Aufmerksamkeitsrichtung stammen. Das Ergebnissignal beinhaltet die vollständige Selektion auf Basis der Lokalisierung einer bestimmten Geräuschquelle aus dem gesamten akustischen Szenenbild.

2 VLSI-Implementierung des neuronalen Netzes

Die Module für die Lokalisierung (Directional s.) und Selektion (Attentional s.) werden mittels neuronaler Strukturen realisiert. Da klassische Modelle neuronaler Netzwerke bisher nicht die Möglichkeit der Behandlung zeitlich dynamischer Signale bieten, wurde hierfür ein stark biologisch motiviertes Netzwerk bestehend aus impulsverarbeitenden Neuronen mit räumlich-zeitlicher Summation gewählt. Aus Gründen der massiven parallelen Berechnungen und der Echtzeitfähigkeit wird an die VLSI-Umsetzung die Forderung nach einer 1:1-Abbildung aller biologischen Netzkomponenten auf einzelne Schaltungsblöcke gestellt.
Das neuronale Netz wurde in Anlehnung an das Modell von Gerstner [GRvH93] entwickelt, es werden jedoch auch einige Elemente des Marburger Ansatzes [ERD90], wie z.B. Informationsaustausch pulscodiert, Nachbildung der After-Hyperbolisation mit Refraktärmechanismus u.a., berücksichtigt. Es basiert auf einem pulspropagierenden Informationsaustausch zwischen einzelnen, weitestgehend analogen Verarbeitungskomponenten (Neuronen, Synapsen). Lediglich die Steuerung und die Schnittstellen der Informationsübertragung zwischen diesen Elementen sind digital ausgeführt, so daß insgesamt von einer gemischt analog-digitalen Struktur gesprochen werden kann.
Für die VLSI-Implementierung des neuronalen Netzes wird eine Standard CMOS Technologie mit Betriebsspannungen ±5V und eine Taktfrequenz von 1MHz benutzt. Zwar reicht für das biologische Modell eine Verarbeitungsgeschwindigkeit im ms-Bereich aus, jedoch wird mit der gewählten Taktrate von 1us die praktische Anwendbarkeit deutlich erweitert. Um die Anwendbarkeit dieses Netzmodelles möglichst universell zu gestalten, werden die Eingangssignale extern mit einem Mikrocontroller aufbereitet. Damit behählt die Chip-Implementierung ihre reguläre Struktur und durch Kaskadierung mehrerer Chips lassen sich umfangreiche Aufgaben lösen.
Eine weitere Forderung an die VLSI-Implementation des neuronalen Netzes ist das permanente Lernen. Als Lernvorschrift wird eine modifizierte Hebb'sche Lernregel zur Realisierung des Spine-Learning benutzt. Die Gewichtsänderung erfolgt laut (1) mit der zeitlichen Korrelation zwischen den Feuergeschichten H_d des Senderneurons und H des Empfängerneurons, wobei diese noch mit einer von aussen einstellbaren globalen Lernrate e multipliziert werden. Auch andere Lernverfahren sind in Abhängigkeit von der konkreten Aufgabenstellung möglich. Dabei ist jedoch zu beachten, dass die Lernvorschrift als Hardware realisiert wird. Somit können zwar Parameter wie die Lernrate von aussen eingestellt werden, das Verfahren selbst ist jedoch fest implementiert.

$$\omega_{ij}^{t} = c \cdot \int_{t-\tau}^{t} o_i^{t-\tau} e^{-\frac{t}{\tau}}\, dt \cdot \int_{t-\tau}^{t} \chi_j^{t-\tau} e^{-\frac{t}{\tau}}\, dt + \omega_{ij}^{t-1} \tag{1}$$

Der Hauptvorteil analoger VLSI-Implementierungen besteht darin, dass sich relativ komplexe arithmetische Operationen mittels weniger einfacher Hardwareelemente realisieren lassen. Dadurch können viele Funktionseinheiten auf einem Chip untergebracht werden. Weiterhin ist die hohe Verarbeitungsgeschwindigkeit analoger Implementierungen zu nennen. Als nachteilig erweisen sich die hohe Rauschempfindlichkeit, Streuungen des Herstellungsprozesses, relativ geringe erreichbare Genauigkeit, sowie Langzeitstabilitätsprobleme der Speicherung analoger Werte.

3 Struktur des neuronalen Netzes

Die Informationsverarbeitung in den Synapse und Neuronen erfolgt ausschließlich lokal. Die Struktur des neuronalen Netzes stellt ein klassisches Array dar. Die Neuronen des Chips bilden einen Vektor, über dem eine Matrix von Synapsen angeornet ist (vgl. Abb. 2). Das ermöglicht eine optimale Verdrahtung, in die nur pulsausbreitende Schnittstellen zwischen Neuronen und Synapsen neben den Steuerungssignalen (Takt, Sägezahn-Spannungen, Lernrate, Referenzspannungen) einbezogen werden. Die gewählte Anordnung erlaubt jede beliebige Netztopologie, von der Vollvernetzung (vgl. Abb. 2) über rückgekoppelte Netze bis zu mehrlagigen Netzen mit Neuronen-Clustering. Die gezielte Unterbrechung einer Verbindung zwischen zwei Neuronen kann entweder durch das Weglassen der zugehörigen Synapse oder durch eine Initialisierung ihres Gewichtes mit dem Wert 0 erfolgen. Im zweiten Fall ist jedoch zu beachten, daß infolge von Lernvorgängen eine Reaktivierung der unterbrochenen Verbindungen möglich ist.

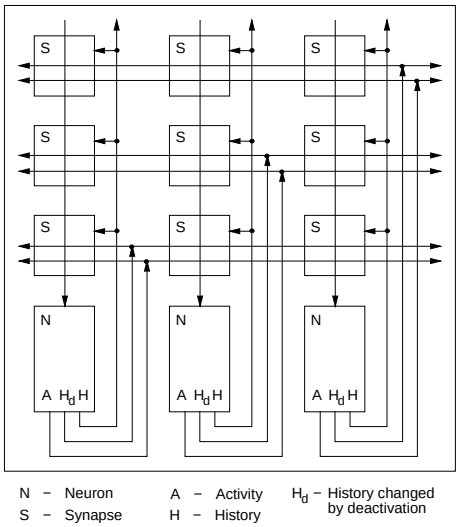

Abbildung 2: Struktur des neuronalen Netzes

Der Informationsaustausch zwischen Neuronen und Synapsen erfolgt über Impulse (Spikes) der Frequenz 1MHz. Wird ein Neuron aktiv, so sendet es einen digitalen Impuls aus. Diese Aktivität A wird an alle Synapsen in einer Zeile verteilt. In den Synapsen wird ein als Spannung gespeicherter Gewichtswert in einen Strom gewandelt und mit der Aktivität des Senderneurons abgetastet. Der Ausgangsstrom wird zum Empfängerneuron übertragen. Die räumliche Summation über alle Stromimpulse der zugehörigen Synapsen einer Spalte wird mittels des Leitungsknotenpunktes am Eingang des Neurons besonders einfach realisert. Der Eingangskondensator im Neuron führt eine zeitliche Summation aus. Mit dieser einfachen Lösung bei der Berechnung des postsynaptischen Potentials im Neuron kann die Netzgröße beliebig gestaltet werden. Lediglich eine Veränderung der Impulsbreite der Ströme ist erforderlich, um die Sättigung des Potential-Kondensators zu verhindern. Da aber die Zahl der implementierbaren Neuronen und Synapsen pro Chip technologisch begrenzt ist, wurde beim Entwurf auch die Möglichkeit einer Kaskadierung mehrerer

Chips vorgesehen.

4 Aufbau des Neurons

Mit der räumlich-zeitlichen Überlagerung im Neuron entsteht am Potentialkondensator Z in Form einer Spannung die postsynaptische Potentialfunktion (EPSP oder IPSP). Dieser innere Zustand des Neurons wird mit einer Schwelle T verglichen. Wird die Schwelle überschritten, so kommt es auf der Aktivitätsleitung zum Aussenden eines zeitlich genau definierten Spikes. Zur Nachbildung der herabgesetzten Feuerwahrscheinlichkeit nach einer Aktivität (After-Hyperbolisation) wird die Schwelle für eine bestimmte Zeit angehoben.
In der Schaltung des Neurons nach Abb. 3 wurde am Eingang ein Abtastglied eingeführt. Die Zeitdauer der abgetasteten Stromimpulse wird in Abhängigkeit von der Netzgröße mit dem Taktsignal Clock2 eingestellt. Zusätzlich erreicht man damit eine Synchronisation aller Synapsensignale. Das Aufladen des Potentialkondensators Z erfolgt linear mit den Stromimpulsen, seine Entladungskurve modelliert die β-Funktion mit einer Ausklingdauer von $30\mu s$. Das Ergebnis des Komparatorvergleichs zwischen der Schwelle und dem inneren Zustand liefert die Neuronaktivität. Diese würde jedoch nach der Anhebung der Schwelle (Refraktär-Mechanismus) innerhalb einer kurzen Verzögerungszeit wieder abfallen, so daß eine Pufferung für die restliche Zeit der Periode erforderlich ist. Dazu wird ein einfacher Speichermechanismus mittels Gate-Kapazität eines Inverters benutzt. Der Kapazitätsspeicher und der Refraktär-Mechanismus werden von zwei gegenphasigen Taktsignalen angesteuert. Dieses Taktregime gewährleistet die Übernahme des richtigen Ergebnisses, die Synchronisation aller Neuronen-Ein- (Clock2) und Ausgänge (Clock1), und verhindert das selbsterregende Aufschwingen eines Neurons.
Im Falle des Feuerns eines Neurons wird sein Refraktär-Mechanismus aktiviert, welches während einer Erholungsperiode für die Dauer der nächsten Takte eine weitere Aktivität verhindert. Dieses biologische Phänomen der After-Hyperbolisation wird mit der Addition eines Impulses zu der Schwelle T modelliert. Durch diese Überlagerung sich ergebende zeitveränderliche Schwelle muß über dem Wertebereich für das postsynaptische Potential liegen, also oberhalb 0V. Man unterscheidet zwischen einer absoluten und relativen Refraktärzeit. Die absolute führt bei jedem Neuron unabhängig von der Anzahl der aktiven Synapsen zur vollständigen Hemmung der Aktivität. Während der relativen Refraktärzeit fällt die Schwellenspannung exponentiell gegen ihren ursprünglichen Wert, so daß der Zeitpunkt einer neuen Neuronenaktivität von der aktuellen Amplitude der Synapsenströme abhängig ist.
Schaltungstechnisch wird der Refraktärmechanismus mit dem Impulsformer und der Addition von Schwellen realisiert. Die Hauptbestandteile des Impulsformers sind ein Kondensator und ein Komparator. Dabei wird die Kondensatorspannung mit einem Wert X verglichen. Im Falle der Neuronenaktivität wird der Kondensator schnell aufgeladen, der Ausgang des Impulsformers ist identisch mit dem H-Pegel des Komparators. Die relative Refraktärzeit beginnt beim Unterschreiten des Wertes X durch die Kondensatorspannung, wobei diese Spannung direkt auf des Ausgang durchgeschaltet wird. Mit X wird die Zeitdauer der absoluten After-Hyperbolisation in den Grenzen 1 bis $3\mu s$ eingestellt, die relative Phase hängt mit der exponentiellen Entladedauer

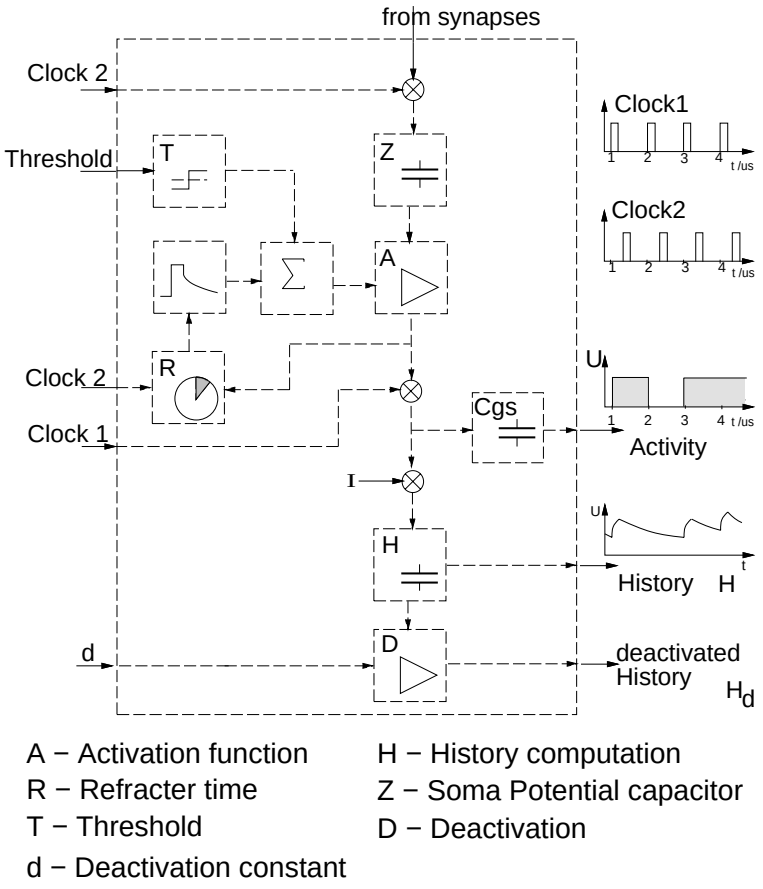

Abbildung 3: Neuron

von $100\mu s$ ($\tau = 20\mu s$) zusammen. Zur Addition mit der Schwelle T kann wahlweise ein Addierer oder ein flächenminimaler Pegelverschieber benutzt werden.
Zusätzlich zur Aktivität werden fortwährend zwei History-Verläufe zur Modifikation der synaptische Gewichte gebildet. Beide werden mit einem Integrationskondensator erzeugt, der im gleichen Zeitregime wie die Eingangsabtastung arbeitet. Dabei ist H die Feuergeschichte des Empfängerneurons (postsynaptische Potential) und H_d die um eine globale Inhibition d verminderte Feuergeschichte des Senderneurons (präsynaptisches Potential). Während jeder Neuronenaktivität wird dieser Kondensator um einen festen Wert ΔU aufgeladen und danach mit $\tau = 6\mu s$ entladen.

5 Synapse

Bei den Synapsenmodellen dieses neuronalen Netzes soll unter lokaler Informationsverarbeitung vorallem eine lokale Gewichtsspeicherung verbunden mit der Lern- und Refresh-Realisierung verstanden werden. Die synaptischen Gewichte werden an einer Polysilizium-Kapazität in Form analoger Spannung gespeichert. Damit liegt das größte Problem bei der Langzeitstabilität des gespeicherten Wertes.

In Abbildung 4 ist die prinzipielle Struktur einer Synapse dargestellt. Die wesentlichen Bestandteile sind die Lernschaltung mit einem Multiplizierer und einer Ladungspumpe, der Gewichtskondensator, seine Refresheinheit und ein UI-Wandler mit Abtastung als Activity-Multiplizierer. Die History-Werte H, H_d zweier Neuronen werden mit einem analogen Gilbert-Multiplizierer verknüpft. Das Multiplikationsergebnis als Differenzsignal wird in einem Spannungs-Zeit-

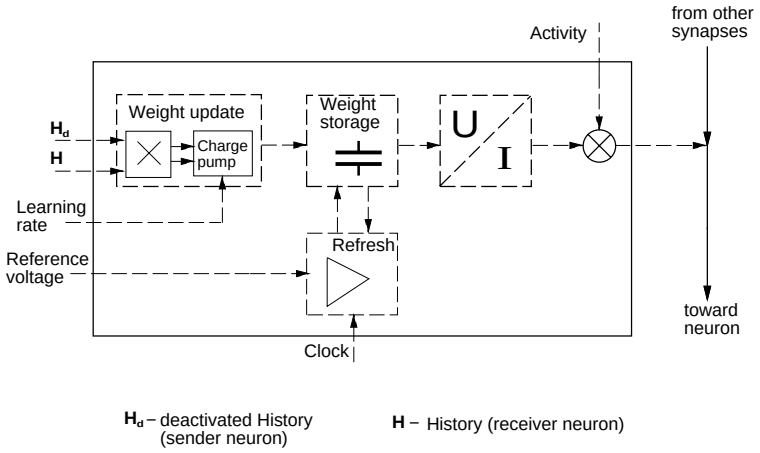

Abbildung 4: Blockstruktur einer Synapse

Konverter in Impulse gewandelt, mit denen ein vorzeichenbehafteter Ladestrom zum Gewichtskondensator zugeschaltet wird. Damit wird eine bestimmte, der Lernvorschrift entsprechende Ladungsmenge auf den Kondensator aufgebracht. In Abb. 4 wurde aus diesem Grund der Schaltungsblock als Ladungspumpe bezeichnet. Die Idee entstammt der Veröffentlichung [MA94]. Der variable, für das ganze Netz global verdrahteter Ladestrom entspricht der Lernrate c aus (1).

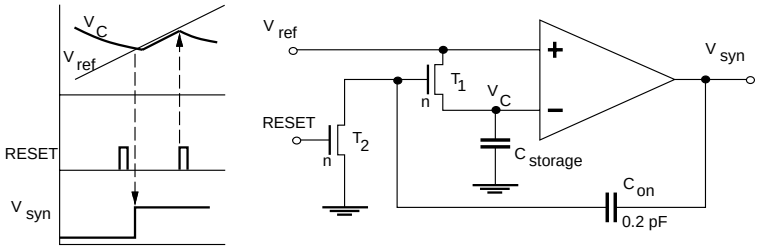

Abbildung 5: Blockstruktur der Refresh-Schaltung

Der single-ended UI-Wandler nach [Wan90] überführt die Kondensatorspannung in einen wertkontinuierlichen Ausgangsstrom, der mit der Senderneuron-Aktivität abgetastet wird. Damit erfolgt die Informationsübertragung von der Synapse zum Empfängerneuron mit non-return-to-zero Signalen (NRZ) einer Frequenz von 1MHz.
Um die Entladung des Gewichtskondensators gezielt gegen das Massepotential zu steuern, wird ein Spannungsteiler, bestehend aus zwei sperrenden Transistoren unterschiedlichen Leitfähigkeit-

styps, eingesetzt. Mit einer Refresh-Schaltung (Abb. 5) wird das Synapsengewicht trotzdem konstant gehalten und der Entladung entgegengewirkt. Der Grundgedanke aus [VOM+91] besteht darin, die Kondensatorspannung ständig mit einer Sägezahn-Referenzspannung zu vergleichen. In dem Moment, in dem die Referenzspannung erstmals unter der Kondensatorspannung liegt, wird der Kondensator der Referenzspannung hinterhergeführt.

6 Zusammenfassung

Es wurde ein System zur Selektion und Analyse von akustischen Geräuschen vorgestellt. Ausgangspunkt bildete dabei die Funktionalität des menschlichen Innenohrs. Bei der Realisierung wurden Methoden und Verschaltungsprinzipien der Neuroinformatik und Biosignalanalyse verwendet.
Die beschriebenen Module der akustischen Signalverarbeitung wurden unter C++ und MATLAB simuliert. Für die Hardwareimplementation sind einzelne Schaltungsblöcke entworfen und für Testzwecke die Synapse in einer $2.4\mu m$ CMOS-Technologie implementiert worden. Mit dem Übergang zur neuen $0.5\mu m$ Technologie werden Testimplementationen des Neurons als auch kleinerer Netze vorbereitet.

Literatur

[ERD90] R. Eckhorn, H. J. Reitboeck, and P. Dicke. Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex. *Neural Computation*, 2(2), 1990.

[GRvH93] W. Gerstner, R. Ritz, and J. L. v. Hemmen. Why spikes? Hebbian learning and retrieval of time resolved excitation patterns. *Biological cybernetics*, 69:503–515, 1993.

[MA94] Takashi Morie and Yoshihito Amemiya. An All-Analog Epandable Neural Network LSI with On-Chip Backpropagation Learning. *IEEE Journal of Solid-State Circuits*, 29(9):1086–1093, 1994.

[VOM+91] E. Vittoz, H. Oguey, M. A. Maher, O. Nys, E. Dijkstra, and M. Chevroulet. Analog Storage of Adjustable Synaptic Weights. In U. Ramacher and U. Rückert, editors, *VLSI Design of Neural Networks*, volume 122 of *The Kluwer international series in engineering and computer science: VLSI, computer architecture and digital signal processing*, pages 47–63. Kluwer Acad. Publ., Boston; Dordrecht; London, 1991.

[Wan90] Z. Wang. *Current Mode Analogue Integrated Circuits and Linearization Technique in CMOS-Technology.* Hartung-Gorre-Verlag, Konstanz, 1990.