A PACED ANALOG SILICON MODEL OF AUDITORY ATTENTION

T.P. ZAHN, R.IZAK, K.TROTT, P.PASCHKE

Technical University of Ilmenau
Dept. of Neuroinformatics and Dept. of Microelectronics
PF 100565, 98684 Ilmenau, Germany

This paper describes a 1MHz CMOS implementation of an integrate and fire neuron network. The system models the human ability to separate unknown sounds under natural conditions. Sound sources are separated based on binaural time delay of auditory nerve spiking pattern. The acoustical attention is guided by the novelty of a sound, acquired knowledge and interaction with other sensoric systems. Signals are coded into current pulses modelling the response of the Inner Hair Cell-Ganglion complex. Uniform paced analog models of pulse propagating cells and synapses are used at different processing levels. A module based layout generator for the realization of different network size and connectivity is under construction. The system is designed to provide a silicon solution for the interaction of acoustical with other sensoric information at the autonomous robot platform MILVA.

1 Introduction

Acting in a natural environment any sensoric driven system is confronted with a overwhelming amount of data. Therefore one of the main questions is the separation of "interesting" information out of the data stream. Like biological systems, the mobile robot MILVA is designed to combine visual, acoustical and tactile sensoric input to determine the focus of attention in an unknown environment. The concept of acoustical attention is mainly concerned with the solution of the object-background problem (Cocktail-Party Effect). Our network approaches this task by reassembling the functionality of the timing pathway in vertebrates for sound source localization. Based on the azimuthal location we extend the WTA model of Lazzaro[3] to a time and knowledge guided selection principle. The implementation of these principles is based on Integrate and Fire (IF) neurons acting as coincidence detectors at several stages of the auditory pathway. The evaluation of timing information at these cells uses binary spikes of fixed pulse width and height with no need for precise calculations. The combination of analog circuitry to model the chemical potentials of neurons and synapses with global digital clocks governing the information exchange, enables the system to ensure timely accurate processing while minimizing the calculational, spatial, and power consumption expense.

2 Functional Structure

The functional structure of our approach consists of four major stages. Each stage roughly summarizes the functionality of a neural structure along the auditory pathway (see Fig.1). But our major concern was not to reassemble the firing behaviour at each of these complex structures. We rather attempt to model the behaviour of some substructures, critical to the task of the robot. The first stage is marked in Fig.1 as inner ear. It physically exists twice assigned to the left and right perception channel. The main parts of our ear are the model of cochlear mechanics and the sensoric coding performed by the inner hair cells.
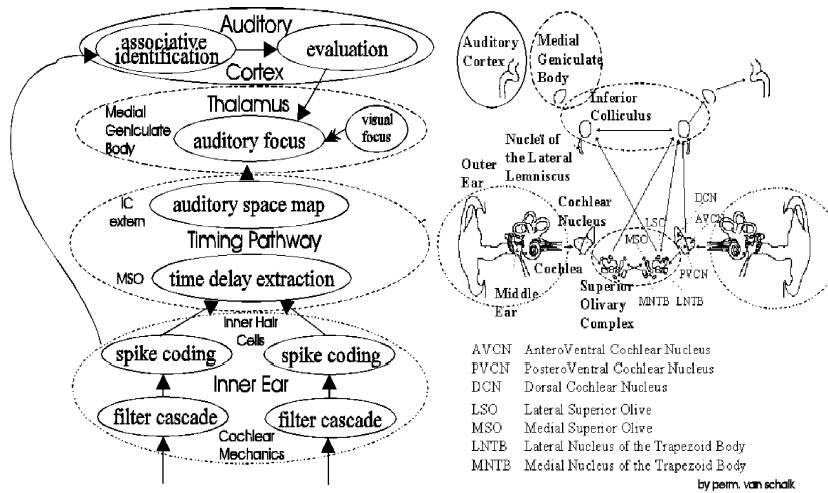

Figure 1: Module structure of the acoustical attention network

The timing pathway accounts for the second stage and uses inter-aural time delay exclusively for the localization of sound sources in space. In the thalamic nuclei at stage three, the guided focusing to one of the detected sound sources is performed. Finally at the cortical level we assume parallel evaluation and associative identification of spike patterns sending efferent attentional signals back to the thalamus.

2.1 Membrane Model

Cochlear mechanics and its simulation in analog hardware has been intensively investigated by the CalTech group of Mead. Before adapting to their findings we tested wavelet based methods, a special kind of ARMA modelling and some combinations of lowpass and resonator filters. Crucial for the adaption to the

Lyon model [5] was its simplicity and the possibility to include outer hair cell functionality for automatic gain control as proposed by Sarpeshkar [8]. Based on second order all pole gammatone filters we use an analog circuit similar to the model of van Schaik [9]. Presenting a single tone stimulus of 1 kHz to a membrane model of a 64 sequential filters, we achieved the response shown in Fig.3. The filter response is duplicating the traveling wave behaviour of the inner ear basilar membrane (top-down) with its maximum elongation at a specific position along this membrane.
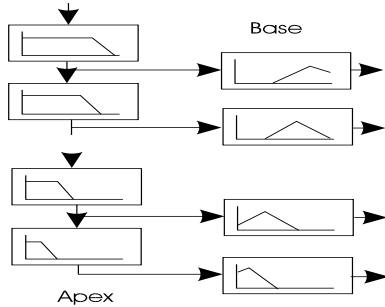

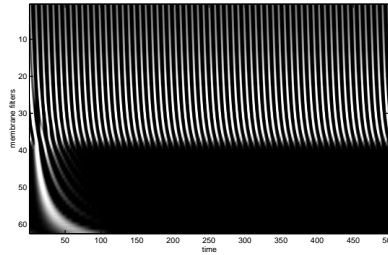Figure 2: Filter cascade as model of the basilar membrane

Figure 3: Response of the basilar membrane model to a 1 kHz stimulus

2.2 Inner Hair Cell Model

The biological coding procedure of the basilar membrane movements into specific timing of spike chains at the acoustical nerve is the functional concern of our inner hair cells. Like most models we omit the duplication principle of nature and assign just one Inner Hair Cell (IHC) to each of the frequency channels preserving tonotopical frequency information this way. Since the stereocilia of an IHC cause depolarizing currents only when moving toward one direction, the resulting soma potential duplicates the movement of the basilar membrane. Comparing this potential with a threshold just above the resting potential leads to the emission of a first spike at a fixed point of the positive movement cycle (phase locking). The number of spikes emitted, depends on the amplitude of elongation and the refractory period. The functional result is the coding of continuous signals into time discrete binary pulse trains. During simulation a stochastic part, likely to occur in analog implementations, was introduced. But stochastic firing of the IHC during absence of input (Fig.4) does not prevent the fibers from strong phase locking at the site of characteristic frequency
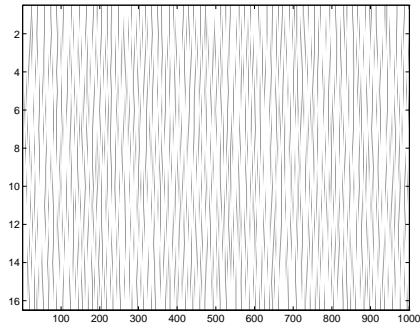
Figure 4: 1000 ms random firing sequence of a 16 channel IHC array during absence of stimulus
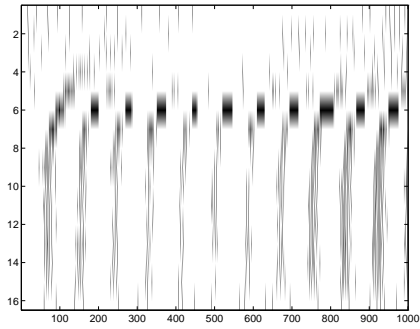

Figure 5: phase locked IHC firing sequence during stimulation by a 1kHz tone

as well as weaker locking at neighbouring locations in response to a sound imposed to the ear. The graded locking and intensity response shown in Fig.5 models the well known tuning curves for acoustical nerve fibers [13].

2.3 Timing Pathway

According to neurophysiological findings and the silicon model of Lazzaro [3] azimuthal localization of sound happens at the site of the Medial Superior Olivar (MSO) nuclei and their projections to the central and external nuclei of the Inferior Colliculus (IC). To model this so called timing pathway, counterpropagating delay lines were used to realize the combination of the left and right hemisphere at the MSO. The principle output of the MSO is the crosscorrelation function between left and right ear calculated separately for each frequency channel. Since information is coded as binary spikes, crosscorrelation can be performed as coincidence detection along delay lines (Fig.6).
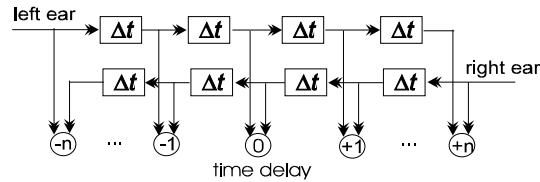

Figure 6: Delay line with coincidence detectors at the MSO

For each of the 16 frequency channels the spike trains of the corresponding frequency channels in the left and the right ear are counterpropagated along a line of 34 delay units. If the spikes enter this structure with a certain interaural delay, they will coincide at a specific location of the delay line and cause firing of the detached detector cell. This way interaural time delay is coded into space. The number of spikes at different locations of the resulting timedelay-frequency map is shown in Fig.7. In response to a complex sound from 20 degree left, it can be observed, that activity (white) of several frequency channels is concentrated around -5 μs delay .
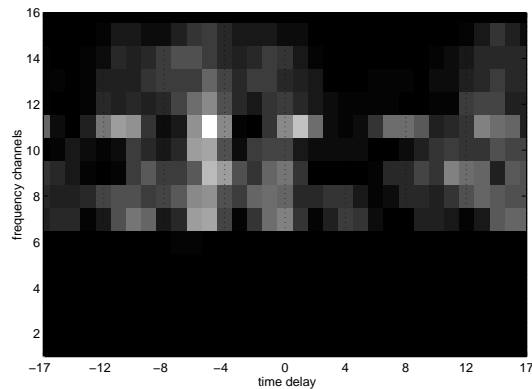

Figure 7: timedelay - frequency map at the MSO where increased spike rates are visible in the frequency channels 7-11 (800-3500 Hz) around $-5\mu s$

The tonotopical organization of isofrequency zones is preserved until the projection from the central to the external nucleus of IC. Here the delay - frequency plane is converted to a time delay map by dendritic summation across frequencies (see Fig.8). Taking advantage of the physically determined
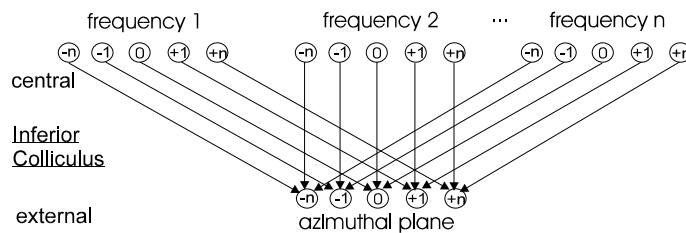

Figure 8: Frequency summation architecture as projection from central to the external Inferior Colliculus

time-delay common to all frequency components of a specific sound source, the azimuthal position of this source relative to the base of the "ears" can be determined. The resulting information is not only cue for the localization of a sound but also separates different sound sources by increased activity at different locations of the azimuthal vector (Fig.9). The spatial resolution of this information depends on the base distance of the ears as well as the coding resolution. A more detailed description of the timing pathway model is to be found in [13].
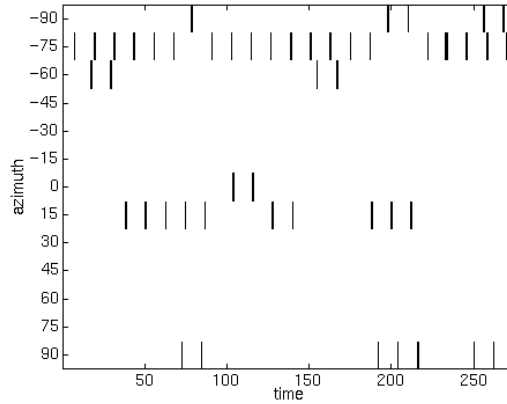

Figure 9: Firing of two nodes in the azimuthal plane at the external IC as result of two sound sources localized 75 degree left and 15 degree right of the center

2.4 Thalamic Auditory Focus

In order to achieve an acoustical focus, one of the detected sound sources has to win over the other activity locations in the auditory space map. The general approach to this problem is the well known Winner Take All network[4] assumed at the Medial Geniculate Body (MGN)of the thalamus. Its output results in a winning unit being the most intense in its firing. Since the intensity of firing in the azimuthal space does not mirror the significance of a sound to the robot, an attentional signal is needed to offset the WTA layer and support the sound source of interest. In order to be significant, a sound has to match two requirements. First, it has to be actually new in the acoustical scene and second it has to be evaluated as "important" by the attentional system. To realize the actuality requirement we installed an activity dependent self inhibition at the MGN. Here an inhibitory connected feed back neuron is assigned to each

focus cell. Receiving the same input, it is suppressing long term firing of the subsequent unit by slowly increasing its synaptic efficiency during each firing. One of the results is, that a moving sound source will achieve full attention until it becomes locally stable and subject to self inhibition. Despite the time dependent suppression method a supporting algorithm is applied to increase the firing probability in case of a positive signal from the attentional system. By coupling the attentional nodes to the focus layer, excitatory firing from the attentional network will lift the soma potential of the focus neurons just below threshold and therefore increase their firing probability. Inhibitory activity from the attentional system will cause the opposite and prevent most focus units from firing at all - the system is "not interested".

2.5   Cortical Evaluation

The task of this module is to develop some knowledge by learning typical sound patterns, which enables the system to exclude them from the attentional procedure. A typical example is the sound of the robot motors. They are detected under varying acoustical conditions but are never of interest to the systems behaviour. The neural structure capable of learning and recalling spatio-temporal spike patterns in an associative manner was first proposed by Gerstner[1]. The fully connected one layer structure employs the implemented Hebbian learning to store a limited number of spike pattern and reconstruct them in case of disturbance by other sources. The dynamic memory receives its input directly from the auditory nerve and is therefore confronted with all of the information in the scene. Any stationary sound will result in a repetitive spike pattern, while instationarities or noise will not cause periodic spiking. The regularity of these pattern leads to a gradual increase of some weights. This results after extensive repetitions at the input in the repetition of the pattern at the output of the memory. For further details see Gerstner[1]. If the learned pattern is part of a complex scene, the memory will only recall the learned pattern and therefore select it from the scene.

To evaluate the significance of regular excitation patterns a hierarchical columnar structure of coincidence detectors is employed to duplicate some features of cortical processing. All channels are fed through a delay line, so the last time steps become available to the mapping procedure at any point in time. Coincidence detectors at the first layer sum over five steps in time and a frequency range of 2 channels. The subsequent layers generalize the spatio-temporal structure to a 3 unit vector at the output. If a repeating pattern contains a regular structure, this will lead to a distinct activity location at the output layer. The similarity to the time delay neurons of Waibel[11] is obvious but the

classification task in our system depends on the judgement whether a sound is significant to the system or not. In the multisensory approach of the MILVA robot this information will be provided by the results of motor actions and visual information. During the current simulations it is provided by the designer and determines if the efferent attentional signal is of excitatory ("interested") or inhibitory ("not interested") nature.

3 Paced Analog-Digital Information Processing

As stated above, we decided to implement the whole network in 2.4 μm analog CMOS technology. The major advantage to us is the fast processing, resulting in the chance to increase time resolution above the natural kHz range to 1 MHz. The complete physically parallel implementation uses pulse propagating IF neurons and distributed hebbian synapses. By combination of these basic cells with modified parameters and different connectivity the specific functionality of the described stages can be achieved. The interconnections within the network are realized by a principle array architecture as displayed in Fig.10. This enables the realization of a wide variety of network topologies, ranging from fully connected toward multi-layer or locally connected architectures. Currently 10 fully connected neurons could be placed on a 50 mm^2 test chip. In case of sparse connections this number increases significantly, but some subnets need even more cells, therefore the chips can be combined to a multi-chip system as shown in Fig.11.
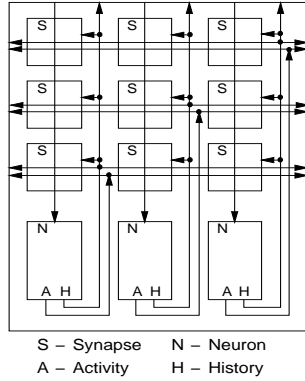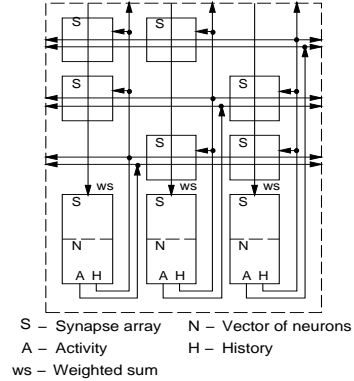

Figure 10: Chip structure   Figure 11: Cascaded chips

Turning toward analog implementations one has to be aware of the inaccuracy and stability problems of such systems. Despite the general robust nature of the processing mentioned above, we used two asynchronous clocks to prevent

the system to a certain extent from oscillations and to ensure time accuracy critical to the function. The third problem of analog systems is the long term storage as well as the accurate modification of synaptic weights. The presented weight refresh and charge pump circuitry is our approach to overcome these limitations.

3.1 Neuron

One of our goals was to find a neural cell model capable to perform at different stages of the system just by tuning of a few parameters. So we could not step as far into biology as van Schaik[9] did, but still attempted to include the essential functionality of integrate and fire cells. The advantage of using a basic cell type, is the preservation of hardware variability and the realization of different neural structures and functions.
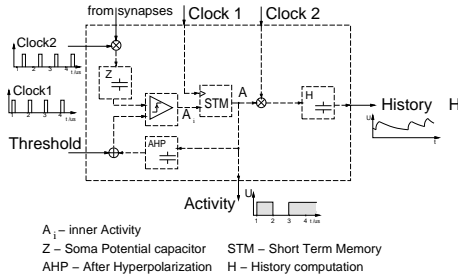
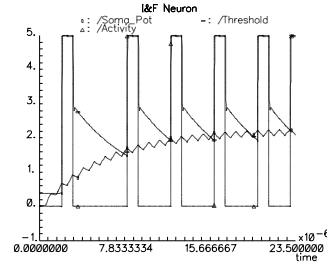Figure 12: Block structure of the neuron model

Figure 13: Simulated internal signals during maximum firing at the dendritic input

The block structure in Fig.12 contains five blocks representing functional circuitry. Each neuron receives spatially added current pulses from affiliated synapses placed in the column above. The analog amplitude of each synaptic current depends on the stored weight and covers the range $\pm 15\,\mu A$. A spatio-temporal summation of all incoming current pulses is realized by charging the Soma Potential Capacitor Z. A discharging resistor in parallel with the $5pF$ capacitor Z approximates the time course of postsynaptic potentials (PSP) as β-function with a fading duration of $30\,\mu s$ and resting potential of $0V$. The neural activity A at the output is the result of the comparison between the soma potential and the threshold. The subsampling of the incoming current pulses by $Clock2$ (1 MHz, pulse width 500 ns) leads to a reduced charge period of the capacitor, extending its functional dynamic range and reducing stabilization

errors at the beginning of the μs interval. If the capacitor voltage Z exceeds threshold, the rail-to-rail comparator generates only a trigger point for the inner activity A_i. To prevent the immediate return due to the refractory process (AHP), A_i is stored in an edge triggered dynamic memory realized as gate-capacitance buffer. The binary activity A at the output of the neuron arises synchronous to the next L-H edge of $Clock1$.(see Fig.13) Each activity pulse is followed by a refractory period. When A turns to high the AHP capacitor is charged and the threshold is lifted to V_{dd}. Therefore the output of the comparator will return to low. This is functionally equivalent to the decrease of soma potential in biological neurons. The AHP potential returns to resting potential in a 2 stage process combining a defined absolute refractory period with an exponential decrease during the relative period. The History (H) circuit is included to model the firing history of the neuron, propagated as dendritic potential back to site of the synapses. During the activity pulse the H capacitor is charged rapidly. Afterwards it is discharged with a time constant of $30\,\mu s$ similar to the Z capacitor. History potential ranges from 2.5 to a maximum of $5V$.
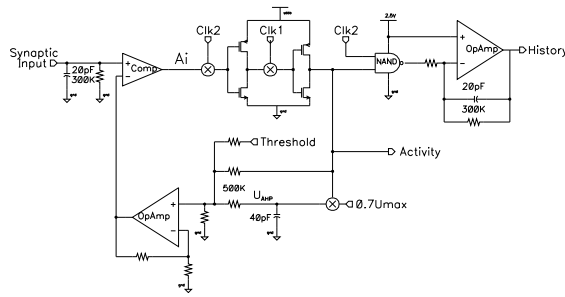

Figure 14: Neuron circuitry

3.2 Synapses with on-chip Hebbian learning

It is widely accepted that the Hebbian learning rule is biological plausible. Although the precise influence of dendritic potentials is not yet clear, Hebbian learning behaviour duplicates best the local process at the synapses. The weight modification will depend on the potential difference resulting from the timing of spikes at the sending and receiving neuron. Enhanced in Fig.15 the weight will increase if the receiving neuron spikes shortly after the arrival of an excitatory pulse. This principle was first proposed by Gerstner [1] for learning in the dynamic associative memory. The synapse is modeled as
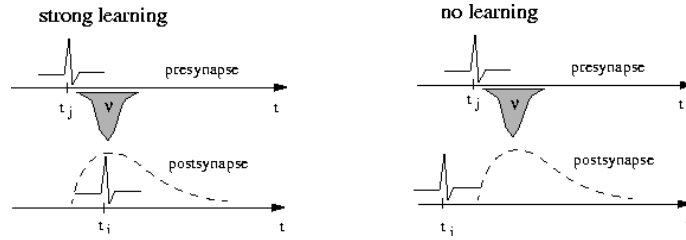
Figure 15: learning principle

the weighted transmission of voltage pulses into current levels at the dendritic tree. A detailed description of the circuits is to be found in Izak et al. [2]. The synaptic weight is stored locally, as a voltage across a $5\,pF$ poly capacitance. Observable in Fig.16, the weight capacitor is charged or discharged by a charge
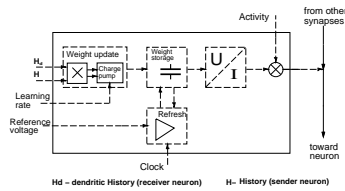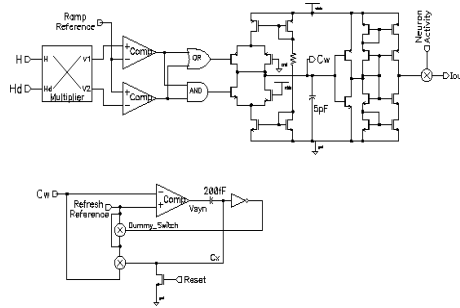

Figure 16: block structure of the synapse

Figure 17: main synaptic circuit

pump depending on the multiplication of dendritic- and postsynaptic potential. In the Gilbert multiplier both history potentials are joined with a linearity of 0.6% at $1V \times 1V$ inputs. The disadvantage of its mV dynamic range can be compensated by higher comparator sensitivity at the subsequent charge pump, based on the idea of Morie [7]. The major problem in analog neural nets is the accurate long term storage of weight values. We defined the necessary accuracy level to 8 bit. By adapting to the idea of Vittoz et al.[10] we use a ramp reference voltage increasing from 0 to 5 V in 256 μs. The voltage across the weight capacitor is continuously compared with this reference. When the reference exceeds the capacitor voltage, the weight is carried along the reference until the next reset pulse occurs. Using a 1MHz reset clock and 256 μs refresh cycle the achieved accuracy fulfills the 8 bit level ($20mV$).

4 Layout Design

Neural hardware design can be divided into all purpose and full custom design chips. To achieve the advantages of full custom design (more detailed modelling, optimal area utilization, etc.) by reducing the overwhelming design expense, we are working on the automation of layout generation. Based on a library of neural elements (cell types, synapses, delay units and others), our design generator is embedded in a CADENCE environment and produces layouts for different net sizes and connectivity structures. The layout of the presented learning synapse is shown in Fig.18.
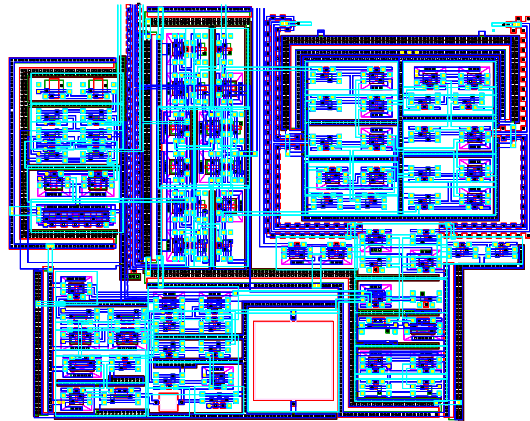

Figure 18: 2,4 μm CMOS layout of a analog synapse with weight storage, refresh and hebbian time resolved learning

Using the 2.4 μm CMOS technology at least 10 Neurons with 100 synapses could be placed on a single chip of 50 mm^2 area. We expect about 20 fully connected neurons to be placed at the same area taking advantage of the upcoming 0.5 μm technology. For larger nets the generator produces partitioned layouts for multi chip modules. Designed by the team of the Graduiertenkolleg in Ilmenau, the generator (Fig 19) will include optimization rules for wire transfer and routing. This way a wide range of networks and functionality becomes accessible to the inexperienced designer.
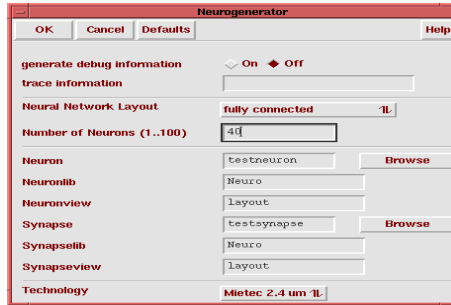
Figure 19: User interface of the layout generator V1.02

5 Conclusion

The simulations and first implementations of neural elements support our postulate, that the combination of analog and digital circuit design is especially well suited to unfold the full power of neural networks. The fully parallel spatio-temporal nature of biological information processing and its simple but often repeated calculations demand a large number of distributed simple processors for high speed and accurate computing. At the same time customized hardware design becomes crucial to the power of applications since neural functionality is mainly based on specific connectivity. Therefore the automation of the design process and the creation of multi purpose neural object libraries is very important to access the power of neural information processing. So far three critical elements, an Integrate and Fire neuron with refractory behaviour and dendritic potentials, a locally learning synapse and an optimized connection have been developed. The synapse is in the second hardware run and the neuron awaits its first realization. Further elements for axonal delays, non learning synapses and specific hair cells are under construction. At different stages of the system we show, that a pulse propagating biological network with very basic elements is capable to perform complex tasks such as separation and identification of sound sources under natural conditions. But the developed hardware-design-system could create visual or cognitive chips as well. To us, preservation of time has been a critical value at every point of the design, since we strongly believe that not just amount but also temporal appearance of biological activity carries essential information how natural systems manage to solve complex problems. Some new theoretical work by Maass[6] and others gives rise to the assumption, that all of the currently used "classical neural networks" can be realized as silicon IF units.

References

1. Gerstner,W., Ritz,R., van Hemmen,J.L. Why spikes? Hebbian learning and retrieval of time-resolved excitation patterns. *Biological Cybernetics* Vol.69 503-115 1993.

2. Izak,R., Trott,K., Zahn,T. Markl,U. Analoge Synapse fuer ein burst-propagierendes Neuronales Netz *presented at the WS on Electronic Integrated Systems E.I.S.* April 1997 University of Hamburg

3. Lazzaro, J.P. and Mead,C. A silicon model of auditory localization *Neural computation* **1** 41-70

4. Lazzaro, J.P. ,Ryckebusch, S., Mahowald,M.A. and Mead,C.A. Winner-Take-All Networks of O(N) Complexity *Technical Report CalTech-CS-TR-21-88* Computer Science Department

5. Lyon,R.F. and Mead,C.A. An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-36, pp. 1119-1134, July 1988.

6. Maass,W. On the Computational Complexity of Networks of Spiking Neurons *TU Graz, Austria Institute of Theoretical Computer Science*

7. Morie,T. Amemiya,Y.: An All-Analog Expandable Neural Network LSI with On-Chip Backpropagation Learning, *Journal of SSC*, Vol.29, No.6, pp.1086 − 1093, 1994

8. Sarpeshkar,R. ,Lyon,R.F. ,Mead,C.A. An analog VLSI cochlea with new transconductance amplifiers and nonlinear gain control. *IEEE Conference on Circuits and Systems*1996

9. A. van Schaik, E. Fragnire, and E. Vittoz, Improved Silicon Cochlea using Compatible Lateral Bipolar Transistors,*in Advances in Neural Information Processing* Systems 8, MIT press, Cambridge MA, 1996, pp. 671-677.

10. Vittoz,E. Oguey,H. Maher,M.A. Nys,O. Dijkstra,E. Chevroulet,M.. Analog Storage of Adjustable Synaptic Weights. *in VLSI Design of Neural Networks,* pages 47-63 Kluwer Acad.Publ.1991.

11. Waibel, A., Hanazawa,T. Hinton,G. Shikano,K., Lang,K.J. Phoneme Recognition Using Time-Delay Neural Networks *IEEE Transactions on Acoustics, Speech and Signal Processing* Vol.37 No.3 March 1989

12. Wang,Z. Current-Mode Analogue Integrated Circuits and Linearization Technique in CMOS-Technology,*Hartung-Gore-Verlag Konstanz, 1990, chapter 6.*

13. Zahn,T.P. Pulse Propagating Network for Attention Based Separation of Acoustucal Signals.*3.Workshop on Bioinformatics and Pulsepropagating Networks* 1996,GfAI ' Berlin