A Mixed-signal Neural Network for Auditory Attention

R. Ižák, T. Zahn, K. Trott, P. Paschke

Dept. Microelectronic Circuits & Systems, Fac. EI,
Dept. Neural Computer Science, Fac. IA,
Technical University of Ilmenau, P.O.Box 10 05 65
98684 Ilmenau, Germany
E-mail: `thomas.zahn@informatik.tu-ilmenau.de`

Abstract

This paper describes a $1MHz$ CMOS implementation of a neural network used for auditory attention. Signals are coded into binary spikes modelling the biology. A uniform model of pulse propagating cell is used in different performing stages. The signal processing is distributed to neurons and synapses realized as analog circuits.

1 Introduction

The concept of attention in an acoustical environment is mainly concerned with the solution of the object-background problem (Cocktail-Party Effect). Our spike propagating network approaches this task by re-assembling the functionality of the timing pathway in vertebrates. Based on the evaluation of interaural time delay of arrival we separate the sound sources in an acoustical scene by their azimuthal location relative to the 25 cm microphone base. A knowledge guided selection principle 14 is employed to extract the most attractive sound source in terms of recent appearance and correspondence with other sensoric systems of the multi sensoric robot MILVA.

2 Functional Structure of the Network

The functional structure of our approach consists of four major stages. The first stage is marked in Fig.1 as preprocessing module. It's main parts are the model of cochlear mechanics and the biological coding performed in the inner hair cell model. To duplicate cochlear mechanics, a sequence of 64 second order all
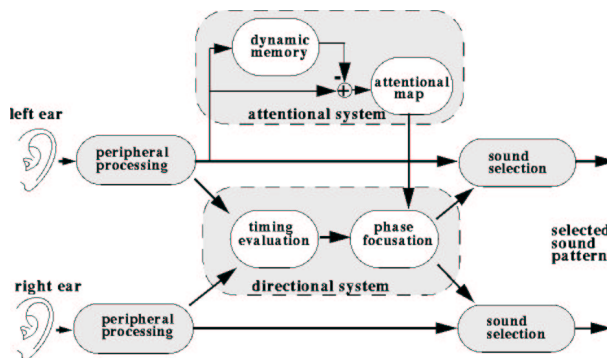

Figure 1: Module Structure of the Acoustical Attention Network

pole gammatone filters [4] has been assigned to each perception channel. The filter response is duplicating the traveling wave behaviour of the inner ear basilar membrane with its maximum elongation at a specific position along this membrane. We achieved correct preservation of inter-aural time delay and computing time was cut down to 1/10 of real time at 32 kHz sampling rate by parallel simulation on a 128 processor CNAPS architecture.

The biological coding procedure of the basilar membrane movements into specific timing of spikes at the acoustical nerve is the functional concern of our inner hair cells (IHC). Like most models [1] we assign just one Inner Hair Cell (IHC) to each of the frequency channels. Since the stereocilla (hairs) of an IHC cause alternating hyper- and depolarizing currents, the resulting soma potential duplicates the movement of the basilar membrane (see Fig.2 upper panel).

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 cy-
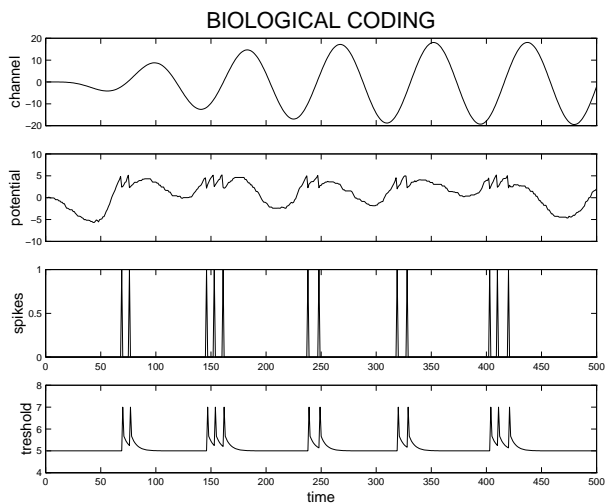
Figure 2: Coding Procedure at a Inner Hair Cell

cle. The number of emitted spikes depends on the amplitude of elongation and duration of the neurons refractory period. The refractory period results from chemical processes in the cell body which prevent the neuron from firing immediately after a spike is emitted and decreases its firing probability during a period of 30 to 100 ms. It has a major impact on the behaviour of the entire system and has been tuned for each IHC separately. All of the acoustical information is now coded into the timing of spikes and will be processed this way throughout the network.

The timing evaluation module is concerned with the extraction of interaural time delays common to all frequency components originating from a specific sound source. It consists of a three layer structure. At the input, delay lines, as in the biological Olivar nuclei, counter-propagate the spikes of corresponding frequency channels from the left and the right ear through 36 neurons in line. Acting as coincidence detectors the cells of the second layer will only fire if spikes at the left and right frequency channel appear with a certain delay, i.e. the component originates from a certain direction. Since many channels are evoked by a sound arriving from a certain direction, the time delay information is several times detected in the time-delay-frequency plane of the second layer. Summing over the frequency channels has been found in biology at the Inferior Colliculus and was introduced in our model by a non learning summation layer at the output of the timing evaluation network. The resulting time delay vector is not only cue for the localization of a sound but also separates different sound sources by increased activity at different locations of the output vector. A more detailed description of the timing pathway model is to be found in [6].

In order to realize an acoustical focus, one of the detected sound sources has to win over the other activity locations in the time-delay vector. The general approach to this problem is the well known Winner Take All (WTA) network [3]. This will result in a winning unit being the most intense in its firing (see Fig. 3).

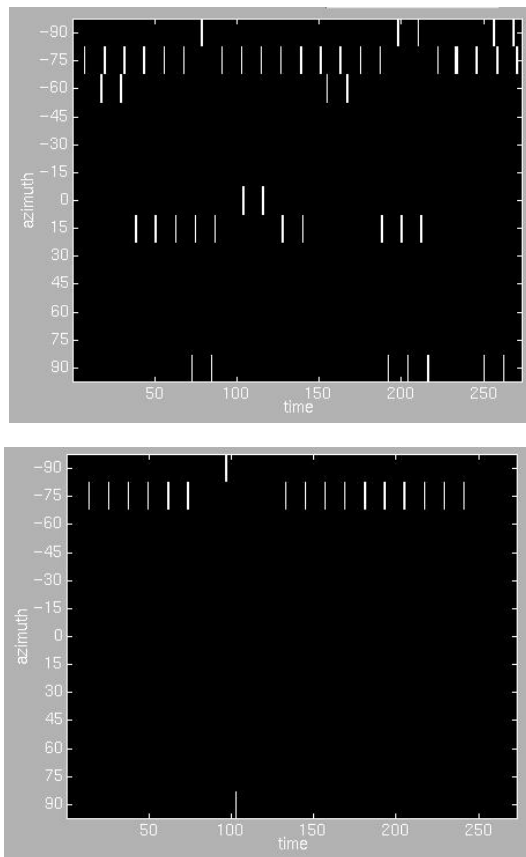Since the intensity of firing in the azimuthal space does

Figure 3: Two flute from different positions - Acoustical Focus on the left flute

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. An "interesting" sound is either a new appearing source or a specific structured signal identified by the attentional map. In order to realize the focus shift to new appearing sounds an inhibitory connected feed back neuron is assigned to each focus cell. It is suppressing long term firing of the winning cell, i.e. the system "gets used to" the sound and returns to global attention. Therefore a new appearing source will immediately take over the focus of attention. 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 atten-

tional 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.

The attentional system contains one layer fully connected dynamic memory as proposed by [7] and a five layer attentional map. The dynamic memory uses Hebbian time resolved learning to store and recall spatio-temporal spike pattern, typical for often repeated but unappealing sounds like the robots motors. These pattern if evoked at the memory input are repeated at the output and subtracted from the original data.

The remaining information is mapped through a five layer structure of frequency-time spike coincidence detectors resulting in a positive or negative attentional signal. It is evaluating the structure of a sound based on the principles of Time-Delay networks. The third information included in the attentional system is the visual focus of attention provided from other systems and introduced as offset to the focus cell of the desired direction.

The sound selection as last stage of the system uses time resolved de-inhibition for the exclusive transmission of spikes with the desired time delay between left and right channel. The resulting information contains the spike pattern of the selected sound in the focus of attention.

3 Hardware Implementation

For the realization of several stages of this system we have chosen a CMOS technology and one power supply of $5V$. Using the $1MHz$ clock rate results in increasing the time resolution above the natural kHz range. The implementation is based on uniform analog pulse propagating cells. Combining these basic cell with modified parameters and different connectivity the specific functionality of different stages can be achieved. Opposite the less accurate processing and stability problems the advantage of such an analog system lies in the less physical layout area and thus allows the complete physical implementation of every neuron and synapse.

The internal information processing of neurons as well as the weight storage of synapses are locally distributed. The interconnections within the network are achieved by choosing an array architecture. This enables simple realization of a wide variety of network topologies. Emphasizing on the evaluation of spike timing a fixed pulse width and height has been chosen. The internal processing is based on two temporal deferred clocks to prevent the system to a certain extent from oscillations and to ensure time accuracy critical to the function.

3.1 Neuron

The neuron cell model was designed with regard to full custom design using library of neural elements. Including the functionality of an integrate and fire cell similar to [1] and capable of variable functionality by tuning its parameters the neuron circuit shown in Fig. 4 was developed. Each neuron receives spatially added
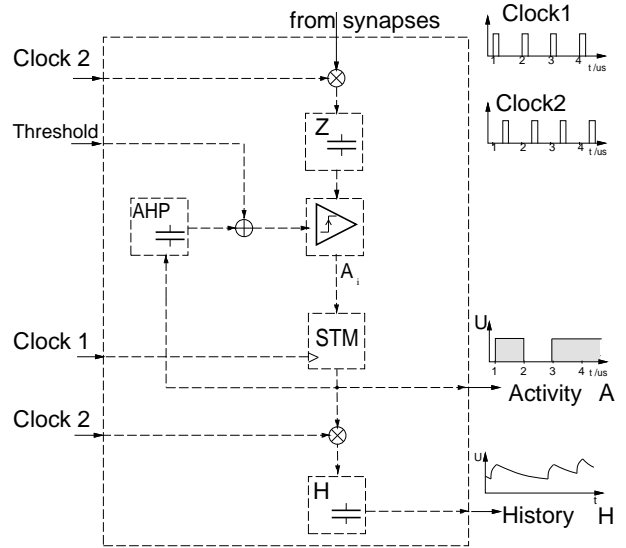
Figure 4: Scheme of a Neuron

current pulses from affiliated synapses placed in the column above. The analog amplitude of each synaptic current depends on its stored weight and covers the range $\pm 15\,\mu A$. The Soma Potential Capacitor Z realises the spatio-temporal sum of all incoming current pulses. The charge of Z accounts for the soma potential of the cell. A discharging resistor in parallel with the $5pF$ capacitor Z approximates the temporal behaviour of postsynaptic potential as β-function with fading duration of $30\,\mu s$ and resting potential of $0V$. The neural activity A is a result of the comparison between the soma potential at Z and the threshold. The determination of activity depends only on the excitatory potential, so the full voltage range from $0V$ to V_{dd} is used for EPSP at Z. The processing controlled by two clocks is divided into an increase of soma potential during the $Clock2$ period and a transmission of activity A synchronous with $Clock1$. If the soma potential exceeds the threshold, the rail-to-rail comparator generates a trigger point for the inner activity A_i. Taking over with $Clock1$ in edge triggered dynamic memory the activity A comes into being. The Short Time Memory (STM) is realized as an gate capacitance buffer.

To reproduce the lower firing probability immediately after a spike is emmited, each activity is followed by

the refractory period. When A turns to high the After Hyperpolarization (AHP) capacitor is charged simultaneously and the threshold is lifted to V_{dd}. The output of the comparator returns to low and therefore the storage of A is essential. The AHP potential returns to resting potential in a two stage process combining a defined absolute refractory period with an exponential decrease during the relative period of $100\,\mu s$.

The History circuit H is included to model the firing history of the neuron needed for local distributed Hebbian learning at the site of synapses. The functionality is similar to the capacitor Z. Increasing the potential by a constant $\triangle U$ during the activity pulse and afterwards discharging with the time constant of $30\,\mu s$. To overcome the current switching the circuit performes by switching voltages between resting potential $(2.5V)$ and $0V$ to an op-amp integrator. Thus history potential ranges from $2.5V$ to maximum of $5V$. The signal is propagated as postsynaptic potential H in the row along with the activity and as dendritic potential H_d to all synapses of the dendritic tree (column).
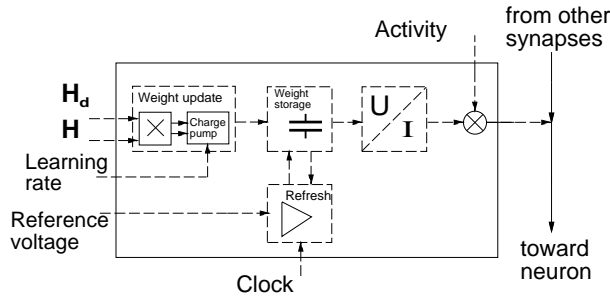
3.2 Synapse

Figure 5: Scheme of a Synapse

The Hebbian learning rule is widely accepted as one of the biological plausible methods. Although evidence for dendritic potential is still missing, its learning behaviour duplicates best the local process at the synapses. The principle first proposed by Gerstner [7] considers the coincidence of sending and receiving neuron potentials. The synaptic weight is increasing only if the receiver spikes shortly after the arrival of an excitatory pulse. Dependent on the task other learning rules are conceivable, but requires fix hardware implementation.

The synapse is modeled as a weighted transmission of voltage pulses into current levels at the dendritic tree. The synaptic weight is stored locally, as a voltage across a $5\,pF$ poly capacitance. Its full voltage range is split into an exciting weight range $(2.5V$ to $5V)$ and into an inhibiting weight range $(0V$ to
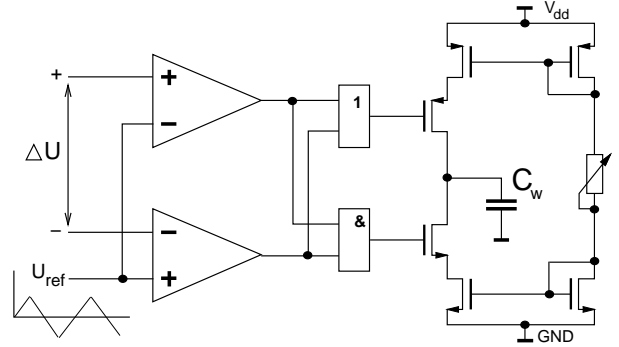

Figure 6: Charge Pump Circuit

$2.5V)$. The modular structure shown in Fig. 5 results from all these requirements. In 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 in the following stage, the charge pump. Based on the idea of [5] the main part, a voltage-to-time converter, consists of two comparators with symmetrical rising and falling slew rates. Comparing the differential multiplier output with a ramp reference proportional voltage pulses are generated. The sign is considered by using two switching lines (AND, OR) as shown in Fig. 6. The weight capacitor C_w is charged with a current switched by these pulses. The capacitor change depends on the histories multiplication as well as the learning rates due to the variable current. Different learning rates can be achieved by modification of the charge current in this stage.

To generate a current output a single input rail-to-rail V-I-converter (see Fig. 7) with saturation in the power supply region is used. The proposal of [8] have been adapted with an inverting amp to attain high input impedance and better linearity. At the output the currentis sampled by activity A of sending neuron and added to other currents in the column line.

The major problem arises with the bad long term storage characteristic of the weight capacitor. A refresh unit (see Fig. 8) similar to the idea of [2] is necessary for the discharge compensation. The voltage across the weight capacitor is continous compared iwth an increasing ramp reference. When the reference exceeds the capacitor voltage, the weight is carried along the reference until the next reset pulse. Using a $1MHz$ reset clock and $256\,\mu s$ refresh cycle the achieved accuracy fulfils the 8 bit level $(20mV)$.
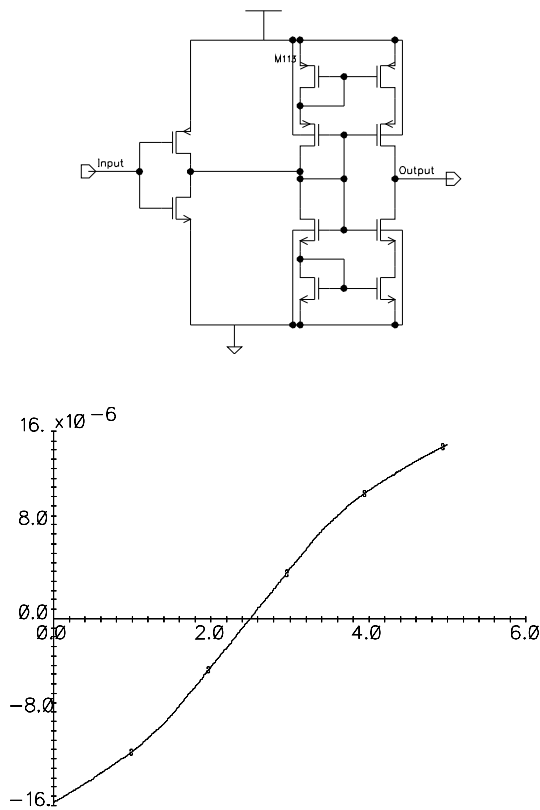
Figure 7: Single Input V-I-converter

4 Conclusion

To make possible variety of implemented neural networks for different applications and still using the performance of full custom design we are simultaneously working on design automation. Embedded in the CADENCE environment the tool kit allows reduction of design expense in the layout generation domain. Based on a library of neural elements it considers different network connections and sizes including partitioned layout for multi chip modules.

An example of the synapse circuit has been implemented in a 2,4μm CMOS technology on a $0.5\,mm^2$ chip area. Due to decreasing the neural connecting
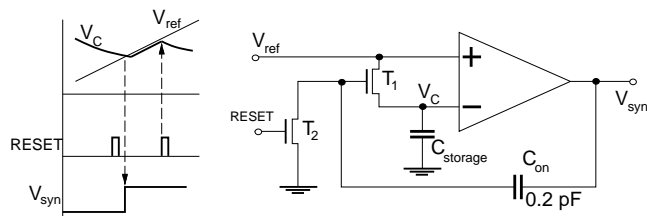

Figure 8: Weight Refresh

complexity at some stages we expect averaged 20 implemented neurons in upcomming $0.5\,\mu m$ technologie on a single chip of $50\,mm^2$.

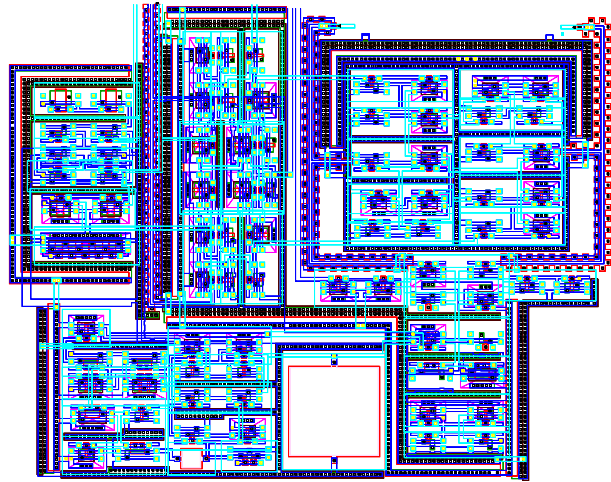This work is a result of the collaboration of depart-

Figure 9: Layout of the Implemented Synapse

ments of Neural Computer Science and Microelectronic Circuits & Systems at the TU Ilmenau and was supported by the DFG.

References

[1] E. Fragnire A. van Schaik and E. Vittoz. *Improved Silicon Cochlea using Compatible Lateral Bipolar Transistors.* Advances in Neural Information Processing Systems 8, MIT press, Cambridge MA, 1996.

[2] M.A.Maher O.Nys E.Dijkstra E.Vittoz, H.Oguey and M.Chevroulet. Analog storage of adjustable synaptic weights. In U.Ruckert U.Ramacher, editor, *VLSI Design of Neural Networks*, pages 47–63. Kluwer Acad. Publ., Boston, 1991.

[3] M.A.Mahowald J.P.Lazzaro, S.Ryckebusch and C.A.Mead. Winner-Take-All Networks of O(N) Complexity. Technical Report CalTech-CS-TR-21-88, Computer Science Department, 1988.

[4] C.A.Mead R.F.Lyon. An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36:1119–1134, 1988.

[5] Y. Amemiya T.Morie. An all-analog expandable neural network LSI with on-chip backpropagation learning. *A Journal of SSC*, 29:1086–1093, 1994.

[6] T.P.Zahn. Pulse propagating network for attention based separation of acoustical signals. *3rd Workshop on Bioinformatics and Pulsepropagating Networks*, GfAI Berlin, 1996.

[7] J.L. van Hemmen W.Gerstner. Spikes or Rates? - Stationary, oscillatory, and spatio-temporal states in an associative network of spiking neurons. *Proc. of ICANN 93*, 1993, pp. 633–638.

[8] Z.Wang. *Current Mode Analogue Integrated Circuits and Linearization Technique in CMOS-Technology.* Hartung-Gorre-Verlag, Konstanz, 1990, chap. 6.