IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2000), Istanbul 2000, pp. II 865-868, IEEE Press 2000.

BINAURAL SOUND LOCALIZATION IN AN ARTIFICIAL NEURAL NETWORK

C. Schauer, T. Zahn, P. Paschke, H.-M. Gross

Dept. of Neuroinformatics, Ilmenau Technical University, D-98684 Ilmenau, Germany

ABSTRACT

We describe a biological inspired model of binaural sound localization using interaural time differences (ITDs). To handle the problem of temporal coding and to facilitate a hardware implementation in analog VLSI, the simulation of the system is based on a spike response model. This neuron model takes up physiological properties like postsynaptic potentials (PSPs) and a refractory period. A winner-take-all (WTA) network selects the dominant source from the representation of the sound's angles of incidences, and can be biased by a multisensory support. We use simulations on real audio data to investigate the function and the practical application of the system.

1. INTRODUCTION

Sound localization is an important function for spatial hearing of human beings and animals. Many investigations and models of auditory perception exist from neurobiology to psychoacoustics [5, 2, 3]. But, although, we could imagine numerous applications in robotics, videoconferencing and speech recognition, only a few working examples are known. When we implemented demonstration software for a mobile robot we noticed one reason for this: the computational demands of digital simulations are extremely high, but special hardware solutions of certain auditory processing tasks are rare. Therefore we strive for a mixed analog-digital VLSI implementation, and use the experiences with software simulations on application-relevant audio data. Our work is related to Lazzaro's neuromorphic auditory localization system [12], but follows a more pragmatic approach. In this paper, we will focus on the architecture and the software simulation of the model.

We assume, that ITD analysis provides a sufficient cue to many localization tasks. In our simulations, we use digital algorithms for the preprocessing and coincidence detection within spike patterns, as well as a uniform spiking neuron in all other parts of the model. One motivation to use spikes is, that a temporal resolution in the range of microseconds is required for the ITD detection. On the other hand, spike patterns can be considered as a consistent way of signal coding which enables a merging of features from different modalities [11].

Figure 1 sketches the system architecture. In the current simulations 16 parallel frequency bands, delivering a spatial resolution of 65 azimuthal angles are computed, whereby the system is simply scalable to practical demands or constraints of the VLSI design. Several stages of the model contribute to localization:

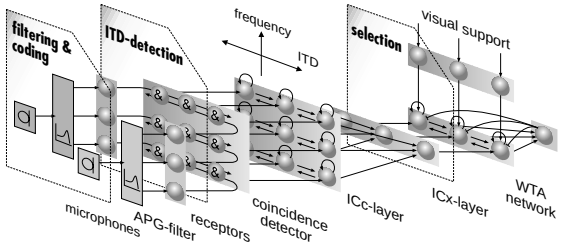

Figure 1: System architecture, composed of the localization model and a following selection mechanism.

1. Filtering and spike coding: The analog signals from two microphones are filtered by a cochlear model (all–pole–gammatone filter) and coded into spike trains (receptors).

2. ITD detection: For every frequency channel the spike patterns from left and right are cross–correlated (coincidence detector). The resulting pattern is stored and postprocessed (ICc layer) and finally projected to a nontonotopic representation (ICx layer) of the azimuthal locations of sound sources in the acoustic scene.

3. Selection: As the result of a WTA process on auditory and visual input, only one direction will be dominant in the final representation.

2. NEURON MODEL

The neuron model (figure 2) is a spike response model inspired by Gerstner's work [6] and takes up fundamental properties of biological cells: the spatial and temporal integration of stimuli via postsynaptic potentials (PSP) in the dendritic tree, the generation of an action potential when reaching a threshold and the effect of diminished sensitivity during a period of refraction. An absolute refractory period and axonal delays are not contained. To describe the impulse response of a synapse, we chose the so called α–function $f_\alpha(t) = \frac{t}{\tau} e^{1-\frac{t}{\tau}}$, the afterhyperpolarization (AHP), which follows a simple exponential fading function. The combination of these potentials results in a biological plausible behavior, which is more complex than the performance of leaky integrate-and-fire models.

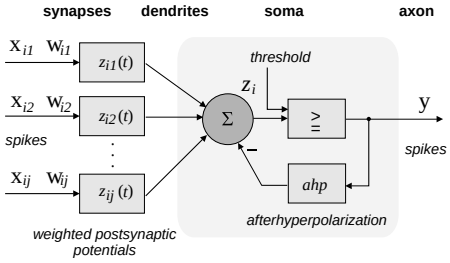

Figure 2: Neuron model

3. COMPONENTS OF THE SYSTEM

3.1. Filtering and spike coding

The preprocessing generates an auditory nerve–like spike pattern from the analog acoustic input signals. The first step is the frequency analysis in the cochlea as the basis for the tonotopic organization of the auditory pathways. Lazzaro's neuromorphic model [12] contains silicon cochleas with 62 output channels to process the analog input directly. In our simulation, the task is solved by a cochlear model using an all–pole gammatone (APG) filter cascade [13]. With respect to the broadband tuning in the auditory nuclei, that are involved in ITD detection [5], we calculate 16 logarithmically arranged channels in the relevant frequency range from 100 Hz to 2.5 kHz from the digitalized microphone signals.

The output of the filter corresponds to the mechanical properties of the cochlear basilar membrane and has to be transformed into a neural response, the specific timing of spike trains in the auditory nerve. This spike coding is realized by a receptor model, simulating the interaction of inner hair cells and ganglion cells. Since their firing is connected with the movement cycles of the basilar membrane, the resulting spike pattern shows the effect of *phase locking* on the acoustic stimulus. The degree of phase locking depends on the refractory properties of the receptor.

3.2. Coincidence detection

If a sound source is not located exactly in the medial-sagittal plane, its position will cause a time difference between the correlated spike patterns in the left and right auditory nerve. According to Jeffress' coincidence model [9] and neurophysiological findings [4], the evaluation of ITD effects is realized by counter propagating axonal delay lines. Coincidence cells, located at different positions along the axons, generate spikes if they receive a simultaneous stimulation from the left and the right hemisphere. Because of the different time delays depending on the length of the propagating fibers, each cell becomes sensitive for a certain ITD. In this way, the temporal information of ITD is transformed into a place code, represented in the spatial distribution of activity in the neural structure.

Numerous extensions have been proposed to the coincidence model of Jeffress, e.g. the suppression of ambiguous responses by a contralateral inhibition [3], the selforganization in the coincidence sensitivity of the cells by Hebbian learning [7] and the usage of bipolar dendrites [1]. In our model a simple abstraction of the function is sufficient – we use a digital delay line and AND gates, which causes a discretization of the angles. Because the maximal delay in the structure must correspond between the model and the real world, the model parameters length N of the delay line and sampling frequency f_s are connected with the physical parameters base distance b of the 'ears' and sonic speed c of the environment by $N = \lfloor f_s \cdot b/c \rfloor$. Using f_s=44.1 kHz, b= 0.25m and c=343 m/s the model can detect $2 \cdot N + 1$=65 directions.

3.3. ICc layer

In the midbrain of birds and mammals ITDs like other auditory features are projected into the Inferior Colliculus (IC), before a further feature extraction and mechanisms of selective attention take place. One possibility to illustrate the feature representation in this auditory nucleus is to describe the formation of maps. These maps of different orientations in the 3-dimensional structure of the central IC display the neural sensitivity to several features, e.g. the tonotopic organization, modulation frequencies, or ITDs [5]. In our model, characteristic frequencies (CF) and characteristic delays (CD) are mapped onto a neural field (figure 3 left). Lateral synaptic connections between ITD-sensitive columns of this field and selfexciting feedback loops are used for a manipulation of the represented feature (figure 3 right). In the opposite to the pure analog and massive parallel processing in the natural auditory system, we have to face jitter effects in the response of the discrete coincidence detector. In the introduced synaptic structure we can prevent those effects and test the influence of certain response properties and coding strategies like spatial smoothing or sharpening of the ITD feature to the postprocessing WTA layer.

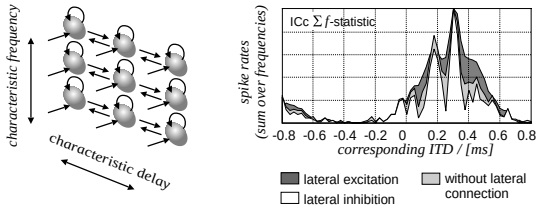

Figure 3: Left: Formation of a 2-dimensional map representing the tonotopy and characteristic delays. Right: The lateral interconnections cause a sharpening or smoothing of the ITD-feature. The response to high frequencies contains periodical components – since the coincidence detection is similar to a cross-correlation of periodical signals, its result is just as periodical.

3.4. ICx layer

In the context of localization, the principle of tonotopy enables to distinguish ITDs from ambiguous phase differences by a recombination of frequency bands. Depending on the band's characteristic frequencies, phase differences are located at different positions in the ITD map. In a conver-

gent projection from many frequency bands, they produce a diffuse activation. The detected ITD position is independent from the tonotopic organization and gives rise to a less ambiguous feature (figure 4). The idea of a summation of the tonotopic response is supported by findings in the IC of the barn owl, where ambiguous activations of single high-frequency bands of the central IC, but a definite response in the nontonotopic extern IC could be observed [12].

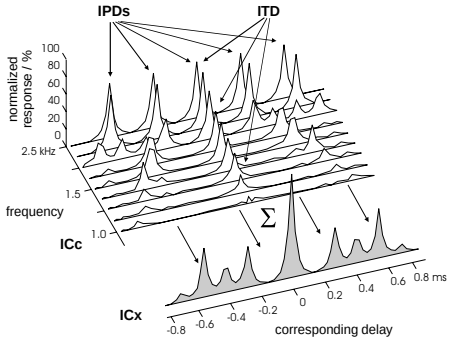

Figure 4: The combination of the tonotopic distributed response of the ICc in a onedimensional ICx-model enables to distinguish ITD from phase differences (IPDs).

While Lazzaro combined the function of the ICx layer and the competition between ITDs in one analog WTA layer, in our experiments with the spike-based WTA and a multisensory input, the separation of the functions to different layers has advantages. The WTA network is stimulated by a maximum spike frequency of just one row of neurons which is independent of the frequency of the acoustic stimulus and limits the dynamic range of the tonotopic distributed feature. This results in a very robust WTA performance.

3.5. WTA selection model

The response of the one-dimensional ICx model often is disarranged by several disturbances like interferences with other sources, echos, or ambiguities which could not yet be suppressed. Modeling aspects of an attentive auditory perception we need to simulate a focusing mechanism, selecting a dominant ITD in the representation of competing features. Findings about cortico-thalamic feedback loops, mechanisms of efferent, inhibitory control and lateral interactions between neurons of the thalamic nuclei [5] suggest the application of winner-take-all (WTA) networks to solve this problem. Our model uses a structure containing lateral and self excitation (like the IC feature map) and an interneuron which integrates the instantaneous activity of the net and generates recurrent inhibition to all cells. In the resulting WTA process only a single region of dominant feature representation can maintain activity [10].

For the application to dynamic acoustic scenes, the selection network should be capable to move the focus of attention to a new sound source. This effect called strong WTA behavior [10] can be achieved by a suitable global inhibition, in particular since we are interested in a decaying WTA activity in the case of silence.

Usually, the attentive perception, especially the localization of objects, has a multimodal character. Various projections from the somatosensory and visual system can be found at the level of the thalamic nuclei [5]. To model an abstract *visual support* to the auditory localization, we first consider where the combination of the two modalities might take place. Since visual features have no interrelations with characteristic frequencies, the first stage for a visual-auditory integration might be the nontonotopic, extern IC. In our system visual information, like the detected skin color of a speaker, is interpreted as the direction of an object of interest. Next we had to decide, whether multiple peaks in the visual inputs are allowed or a single location as the result of a preceding selection process is required. In the sense of an efferent support, only one direction is supported at a time. The actual inclusion of the support is finally realized via additional inputs to the WTA neurons. This is a very simple approach, because it assumes, that the auditory and visual coordinates are already aligned.

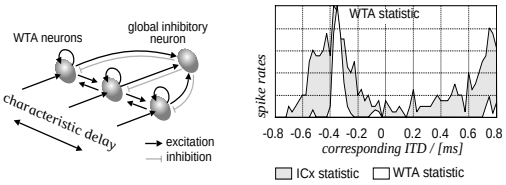

Figure 5: Left: Structure of the WTA network. Right: Result of the selection process.

4. SIMULATIONS AND RESULTS

The localization system was tested offline with data recorded in an open environment including background noise but only little echo effects. Narrow and broadband sounds, including numerous speech signals, were recorded.

The localization of single sources proved to be robust – all directions of broadband sounds were determined correctly. Figure 6 illustrates the focusing to a moving source, emitting pink noise. While the ICx layer displays a diffuse activation and disturbances, the capability of the WTA network to detect a dominant ITD leads to a clear feature representation. The focus stays stable, even if the sound source is moving, which is an important feature of the strong WTA dynamics (figure 6).

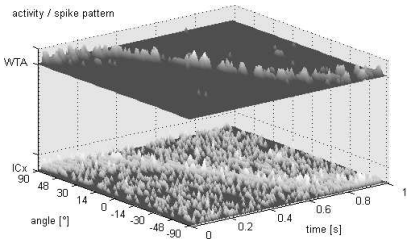

Figure 6: Model behavior for a moving source emitting pink noise. Visualization of the ICx output (bottom) and the activity of the WTA neurons (top).

If multiple sources are present in the acoustical scene, the requirements to the localization system change considerably. Because of interferences between periodical sound components, the dominance of a certain source has to be caused by its intensity or broader spectral constitution. The experiment shown in figure 7 demonstrates, how the focus of attention is shifted from a narrowband sound toward a voice stimulus.

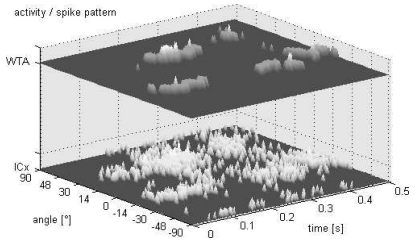

Figure 7: Localization of a narrow band signal and a human call setting in after 100 ms.

Finally, the effect of simulated external support to the WTA process is shown. The system is able to bridge short breaks in the acoustic signal or keep the focus on a non-dominant source (figure 8).

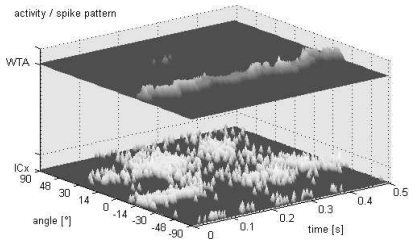

Figure 8: Repetition of the previous experiment, with external support to the narrow-band signal.

In additional indoor experiments we noticed, that the WTA process is able to focus on a sound in about 10ms – often unaffected by the first echos reaching the microphones. For most broadband signals, this time is longer then the arrival of a first wavefront, which has been considered as the longest part of reverberate signals one can localize. But only if a voiced sound hit the room's resonance frequency (in our recordings resonances build up after 30ms), the focus of the WTA layer may be shifted apperently to a random position of an interference (figure 9). This way, although it was not our intention, we can model major aspects of the precedence effect – the dominance of the original sound event up against it's echos. Adding a simple onset detector and with the constraint of a comparatively quiet enviroment, the introduced model is a suitable tool to localize command words under reverberate conditions.

The experiences of the software simulations crucial influenced the design process of a mixed analog-digital hardware system, which VLSI implementation is currently in progress. In particular the usage of real world data and

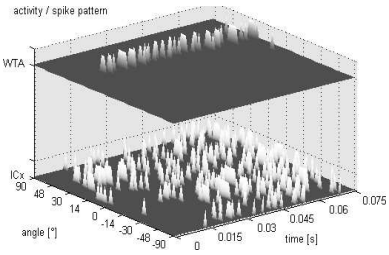

Figure 9: Onset selection in reverberate environment.

tests on robot systems are important to understand the problem and the model in details hidden to purely theoretical investigations.

REFERENCES

[1] Hagai Agmon-Snir, Catherine E. Carr, and John Rinzel. The role of dendrites in auditory coincidence detection. *Nature*, 393:268–272, May 1998.

[2] Yehuda Albeck. Sound localization and binaural processing. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*,891-895.MIT Press,1995.

[3] Jens Blauert. *Spatial Hearing : The Psychophysics of Human Sound Localization*. MIT Press, 1996.

[4] C.E. Carr and M. Konishi. A circuit for detection of interaural time differences in the brain stem of the barn owl. *Journal of Neuroscience*, 10:3227–3246, 1990.

[5] G. Ehret and R. Romand, editors. *The Central Auditory System*. Oxford University Press, 1997.

[6] Wulfram Gerstner. *Kodierung und Signalübertragung in neuronalen Systemen*, vol. 15 of *Reihe Physik*. Verlag Harri Deutsch, Thun - Frankfurt a. M., 1993.

[7] Wulfram Gerstner, Richard Kempter, J. Leo van Hemmen, and Hermann Wagner. A neuronal learning rule for sub-millisecond temporal coding. *Nature*, 383:76–78, September 1996.

[8] R. Ižák, G. Scarbata and P. Paschke. Sound source localization with an integrate-and-fire neural system. In *Proc. of 7th International Conference on Microelectronics for Neural, Fuzzy, and Bio-Inspired Systems MicroNeuro'99*, pages 103–109, Granada, Spain, April 1999. IEEE Computer Society.

[9] L.A. Jeffress. A place theory of sound localization. *J. of Comp. Physiol. Psychol.*, 41:35–39, 1948.

[10] Samuel Kaski and Teuvo Kohonen. Winner-Take-All Networks for Physiological Models of Competitive Learning. *Neural Network*, 7(6/7):973–984, 1994.

[11] Susumu Kuroyanagi, Akira Iwata, *Auditory Pulse Neural Network Model to Extract the Interaural Time and Level Difference for Sound Localization. In* IEICE Transaction on Information and Systems,*E77-D,1994.*

[12] John Lazzaro, Carver Mead. A silicon model of auditory localization. *Neural Computation*,1(1):41-70,1989.

[13] Malcolm Slaney. Lyon's cochlea model. Technical Report 13, Apple, Advanced Technology Group, 1988.