Intern. Conf. on Art. Neural Networks (ICANN), Vienna, 2001, pp. 1141-1146, Lecture Notes in Computer Science vol. 2130, Springer Verlag 2001

A Model of Horizontal 360° Object Localization based on Binaural Hearing and Monocular Vision

C. Schauer, H.-M. Gross

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

Abstract. We introduce a biologically inspired localization system, based on a "two-microphone and one camera" configuration. Our aim is to perform a robust, multimodal 360° detection of objects, in particular humans, in the horizontal plane. In our approach, we consider neurophysiological findings to discuss the biological plausibility of the coding and extraction of spatial features, but also meet the demands and constraints of a practical application in the field of human-robot interaction. Presently we are able to demonstrate a model of binaural sound localization using Interaural Time Differences for the left/right detection and spectrum-based features to discriminate between in front and behind. The objective of the current work is the multimodal integration of different types of vision systems. In this paper, we summmarize the experiences with the design and use of the auditory model and suggest a new model concept for the audio-visual integration of spatial localization hypotheses.

1 Introduction

In recent years a lot of promising work on the problem of spatial hearing has been published – many investigations and models of auditory perception exist from neurobiology to psychoacoustics [2, 1]. However, although numerous applications in robotics and human-machine interaction are imaginable, only a few working examples are known. There could be different reasons for that: on the one hand, the models normally can include only a few details of the complex neural coding and processing mechanisms in the real auditory system. On the other hand, when aiming at localization systems working in everyday environments, many acoustic effects arising from very different acoustic characteristics must be faced.

Moreover it is surprising, that there are still hardly any multimodal approaches, even though artificial vision systems provide processing of motion, color or any object specific feature and the mechanisms of spatial hearing and vision complement one another quite obviously (see table). A multimodal approach is both self-evident following the auditory modeling, and promising advantages in the orientation behavior of mobile robots [3], used to demonstrate the models. Furthermore, some remarkable publications on the neurophysiological background of multisensory integration [8], [9] inspire new solutions for our computational models.

spatial vision	**spatial hearing**
topologically organized receptor fields	due to purely temporal receptor, spatial information need further processing (computational map)
activity in topological maps not necessarily corresponds to objects	activity in computational maps mostly corresponds to objects
time–continuous representation	time–discontinuous representation
limited topological range of receptor	complete spatial mapping

2 Modeling Binaural Sound Localization

In contrast to visual perception, hearing starts with one–dimensional, temporal signals, whose phasing and spectrum are essential for the localization. To evaluate spatial information of a sonic field, the auditory system utilizes acoustic effects caused by a varying distance between the sound source and the two ears and the shape of the head and body. We can categorize these effects in intensity differences and time delays. In [1] a comprehensive study of sound localization based on different types of binaural and monaural information is presented, including findings about the localization blur: The achieved precision in the horizontal plane corresponds conspicuously to the relation of azimuth angle variation and interaural time differences (ITDs) – a hint for the importance of ITD processing. The assumption, that many localization tasks could be solved just by calculating ITDs and the detailed functional and structural description of ITD processing neural circuits has been the starting point of our modeling.

2.1 Binaural Model Concept

Our work on real–world–capable ITD processing is similar to Lazzaro's neuromorphic auditory localization system [5], but follows a more pragmatic approach. In our simulations, we use digital algorithms for the preprocessing and coincidence detection within spike patterns, as well as an uniform spike–response neuron in the other parts of the model [6]. The motivation to use spikes is, that ITD detection requires a timing that is more precise than the description of neural activities by firing-rates. Moreover, spike patterns can be considered a consistent way of signal coding, which enables a merging of features from different modalities. Figure 1 sketches the architecture, including an extension for a simple in front/behind discrimination. The stages of the model:

1. Microphone signals are filtered by a cochlear model (all–pole–gammatone filter) and coded into spikes (hair-cell model).
2. For every frequency channel, the spike patterns from left and right are cross–correlated (Jeffress coincidence detection model for the medial superior olive (MSO) [4]) - the time–code of binaural delay is transformed into a place code representing interaural phase differences.

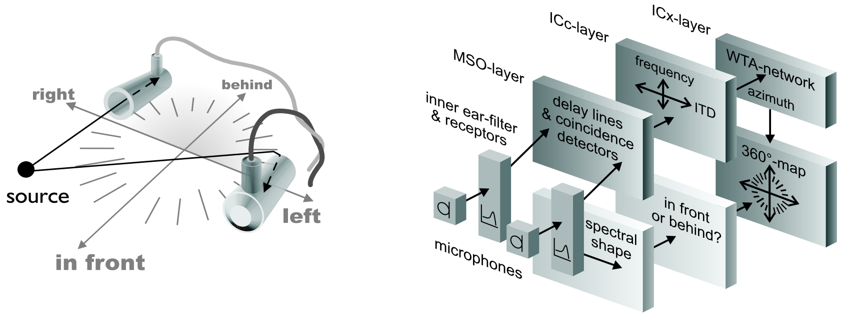

Fig. 1. Left: Microphone configuration: The opposite aligned tubes produce binaural spectral differences, whereby ITDs are not significantly affected. Right: System architecture, composed of the biologically inspired model of ITD detection (dark blocks), a spectrum–based in front/behind discrimination (light blocks). A more detailed description of the ITD-based system is published in [6].

3. The resulting pattern is projected onto a non–tonotopic representation of ITDs and thus of azimuthal locations of sound sources (Model of the Inferior Colliculus, IC). As the result of a winner-take-all (WTA) process, only one direction will be dominant at a time.
4. With the help of a special microphone configuration (see figure 1 left), a simple estimation of interaural spectral differences determines the in front or behind orientation. A 360°-map of horizontal directions is formed.

2.2 Performance of the Sound Localization

Performance tests included all sorts of common sounds (clicks, hand claps, voices, pink noise) and were performed outdoors (without echos) and in an empty, acoustically disadvantageous (echoic) lecture hall. In quiet situations (background noise $<$ -30dB), 100% of the test signals were localized correctly within the acuracy of the discrete system. In additional tests in a shoping center (less echoic, signal-noise ratio 3-5dB) command–words and hand–claps of a person were detected with a probability of 81% and a precission of +/- 10° (90% within +/- 20°). To demonstrate the ability of detecting even moving natural stimuli, the processing of a 12 word long sentence is shown (figures 2), where the tracked speaker position is traveling once around the microphones (performed in a lecture hall without background noise). The experiment demonstrates the superiority of the proposed model over conventional correlation methods that easily fail to process voiced sounds in reverberant conditions. The special properties from which our model benefits can be formulated as follows:

– The spike-pattern mainly codes temporal information like phasing - amplitude is coded indirectly by the spike rate - the effect is a sharp peak as correlation result instead of a smooth sine–response.

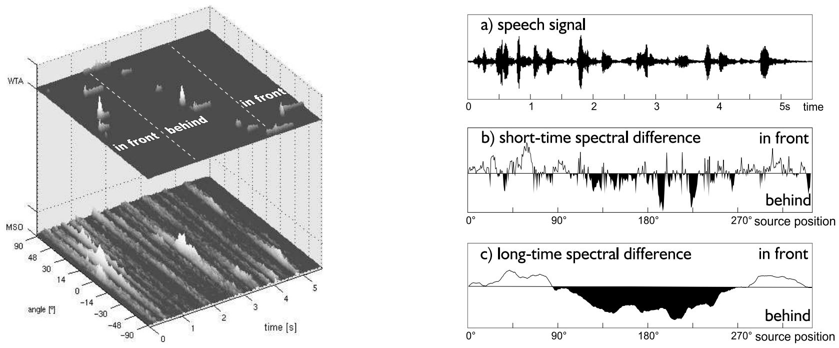

Fig. 2. Left: 360-degree localization of a speaker moving around as described in the text. Visualization of the IC output (bottom) and the resulting activity of the WTA neurons (top). Right: a) Speech–signal, b)short–term (12.5 ms) and c) long–term (0.5 s) spectral difference.

- The broadband response of the cochlea filter together with the tonotopically distributed processing and the recombination of frequency bands effectively prevents ambiguous correlation responses.
- Because competing sound sources, noise or the disturbance by interferences need energy and time to shift the focus in the WTA–layer, the system shows a hysteresis property and thus prefers onsets.

Based on these qualities the system integrates mechanisms of onset, transient, and ongoing sound processing, and realizes the localization beyond the first wavefront. Due to the hysteresis of the WTA-layer, we become adaptive to the overall loudness and sensitive to onsets or changes.

3 Concept for a Sub-Cortical Auditory-Visual Integration

Usually, the attentive perception, especially the localization of objects, has multimodal character. Spatial hearing and vision seem to be complementary strategies in localization. Other than the auditory system, vision is based on a receptor, that is already providing topologically organized information and the question becomes in which way objects of interest manifest themselves in the continuous visual representation? In the context of human-robot interaction, we have named feature candidates like pure intensity, motion, color or contour. In contrast to the low-level auditory-space processing in the midbrain we must now distinguish between cortical object recognition and low-level multi-sensor fusion. Firstly we clarify the term "low-level": Since visual features have no interrelations with characteristic frequencies, the first stage for a visual-auditory integration can be found, following the projections from the non-tonotopical spatial maps in the extern IC. Investigations on the mainly visually, but also auditory (via ICx) innervated, Superior Colliculus (SC), provide evidence for a merging of the sensor modalities and the forming of multisensory spatial maps [8]. Visually sensitive neurons found here, are not or less specialized for color or orientation of contours but respond to broadband intensity and certain velocities of moving stimuli (changes in intensity). We use these findings as a basis for our multi-modal model, although we also consider to integrate higher-level feature as an option in applications. According to [9], at least the following properties of the representation and integration of multiple sensory inputs in the SC are to be considered in the model architecture:

Superficial SC (SCs) is responsive only to visual stimuli. Counterpart of a retinotopically ordered map in SCs is a one-dimensional map of horizontally arranged intensity or intensity differences, provided by a wide-angle vision system.

(i) Deep layers of the SC (SCd) respond to visual, somatosensory, auditory and multi-modal stimuli. (ii) visual receptive fields (RF) are significantly larger than in SCs. In the model we propose convergent visual projections from SCs to SCd, where also the auditory input from ICx is received. According to the field of vision, the RFs of the visual projections cover just a part of the resulting multisensory map.

Most SCd multisensory neurons display response enhancement when receiving spatially and temporally coincident stimuli but show response depression if simultaneous stimuli are spatially separated. This actual property of multisensory integration can be realized by a WTA-type network with both auditory and visual afferents and global inhibition. Competing features inhibit each other, aligned stimuli excite one another.

In SCd overlapping multisensory and motor-maps initiate overt behavior. We are going to use the multisensory map to code turn reflexes of a robotic head toward the acoustic or visual stimulus, small moves if the stimuli originate almost from the center, and stronger ones if "something" is to be seen on or heard from the side.

Different modality-specific RFs have to be aligned to allow response enhancement, even if eyes and/or ears can be moved separately. If so, there has to be also an exclusively visual map in SCd, controlling eye movement. This is consistent with the known models of saccade generation [7]. To achieve map alignment every eye-specific motor-command must cause a shift in the auditory map (in the model the weights in a variable ICx–SCd projection change).

4 First Results and Outlook

The implementation of the model as seen in figure 3a) is currently in progress and hardly depends on a realtime capable simulation and the integration of an experimental robotic platform. This will be the supposition for interactive map shifting for RF alignment or tests in real man-machine communication, but we can already demonstrate the behavior of the WTA-implementation of the multisensory feature map based on offline recorded data, Figure 3 b).

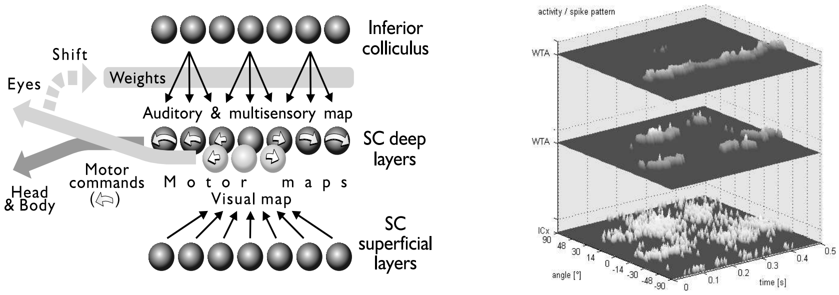

Fig. 3. Left: A simplified and universal model of the superior colliculus, which satisfies all properties, mentioned in the text. Further simplifications can be made, if no separate camera turns are possible (no separate SCd visual map and static ICx-SCd-projections) or if an omnidirectional camera is used (modified SCs-SCd projection). Right: Localization of a narrow band sound (whistle) and a human call setting in about 100 ms later. Response of the auditory model (middle), and the multisensory map (above). The dynamic range of the (spike-based) multisensory map enables local response enhancement and focusing on the multimodal event (an artificial visual stimulus in the form of a lamp-position was applied).

The practical aspect of the application of the model is an expected significant advantage in the detection and tracking of users by the multimodal robot. Besides, we will have to discuss in detail the role of sub-cortical multisensory integration in attentive perception or in control of overt behavior. The interaction in real man-machine communication surely will help to gain insights that are hidden to purely theoretically investigations.

References

1. Jens Blauert. *Spatial Hearing : The Psychophysics of Human Sound Localization.* MIT Press, 1996.
2. G. Ehret and R.Romand The Central Auditory System. Oxford University Press, New York 1997.
3. Gross, H.-M., Boehme, H.-J. PERSES - a Vision-based Interactive Mobile Shopping Assistant. in: Proc. IEEE SMC 2000, pp. 80-85.
4. L.A. Jeffress. A place theory of sound localization. *J. Comp. Physiol. Psychol.*, 41:35–39, 1948.
5. John Lazzaro and Carver Mead. A silicon model of auditory localization. Neural Computation, 1(1):41–70, 1989.
6. Schauer, C., Zahn, Th., Paschke, P., Gross, H.-M. Binaural Sound Localization in an Artificial Neural Network. in: Proc. IEEE ICASSP 2000, pp. II 865-868.
7. P.H. Schiller A model for the generation of visually guided saccadic eye movements. in: Models of the visual cortex, D. Rose, V.G. Dobson (Eds). Wiley, 1985, pp 62-70
8. Stein, B.E. and Meredith, M.A. The Merging of the Senses. The MIT Press, Cambridge, Massachusetts
9. M. T. Wallace, L. K. Wilkinson, B. E. Stein. Representation and Integration of multisensory Inputs in Primate Superior Colliculus. Journal of Neurophysiology. Vol. 76, No.2: 1246-1266, 1996