Proc. of the 25th Pattern Recognition Symposium (DAGM 2003), Magdeburg, pp. 362-369, Lecture Notes in Computer Science 2781, Springer Verlag 2003

# A Computational Model of Early Auditory–Visual Integration

C. Schauer, H.-M. Gross

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

**Abstract.** We introduce a computational model of sensor fusion based on the topographic representations of a ”two-microphone and one camera” configuration. Our aim is to perform a robust multimodal attention-mechanism in artificial systems. 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 applications in the field of human-robot interaction. In contrast to the common technique of processing different modalities separately and finally combine multiple localization hypotheses, we integrate auditory and visual data on an early level. This can be considered as focusing the attention or controlling the gaze onto salient objects. Our computational model is inspired by findings about the inferior colliculus in the auditory pathway and the visual and multimodal sections of the superior colliculus. Accordingly it includes: a) an auditory map, based on interaural time delays, b) a visual map, based on spatio-temporal intensity difference and c) a bimodal map where multisensory response enhancement is performed and motor-commands can be derived. Along with our experiments, questions arise about the spatial and temporal nature of audio-visual information: Which precision or what shape of receptive fields are suitable for grouping different types of multimodal events? What are useful time windows for multisensory interaction? These questions are rarely discussed in the context of computer vision and sound localization, but seem to be essential for the understanding of multimodal perception and the design of appropriate models.

## 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 [3, 2]. However, although numerous applications in robotics and human-machine interaction are imaginable, only a few working examples are known. There might 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.

In computer-vision the situation is different. The field is established and a huge number of models and applications exists - biologically motivated approaches or technical solutions of specific application problems. It is surprising, that multimodal approaches are relatively seldom, even though artificial vision systems provide processing of motion, color or other object specific features and the mechanisms of spatial hearing and vision complement one another quite obviously. For us, the simulation of early auditory-visual integration is promising significant advantages in the orientation behavior of mobile robots [4]. Furthermore, some remarkable publications on the neurophysiological background of multisensory integration [10], [11] inspire new solutions for computational models.

Parts of the model described here, are comparable to the system by Rucci, Edelmann and Wray [7], because a direct structural realization of neural mechanisms is used instead of abstract statistical methods. In contrast to Rucci's system, the emphasis of our work is not placed on the problem of self-calibration and adaption but on robustness and real-world capability. For this reason, also the feasibility of significant and reproducible experiments is discussed in this article.

## 2 Modelling 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 [2] 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 intensity 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 the 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 [6], but follows a more pragmatic approach. In our simulations, we use digital algorithms for the preprocessing and coincidence detection within the auditory patterns, as well as an Amari-type dynamic neural field for the evaluation of ambiguous localization hypothesis [8]. The model includes the following stages:

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) [5]) - the time–code of binaural delay is transformed into a place code, representing interaural phase differences.
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 fig. 1, left), a simple estimation of interaural spectral differences determines the in front or behind orientation. This way, 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 echoes) 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 accuracy of the discrete system. In additional tests in a shopping 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 precision 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 1), where the tracked speaker position is travelling once around the microphones (performed in the lecture hall without background noise).

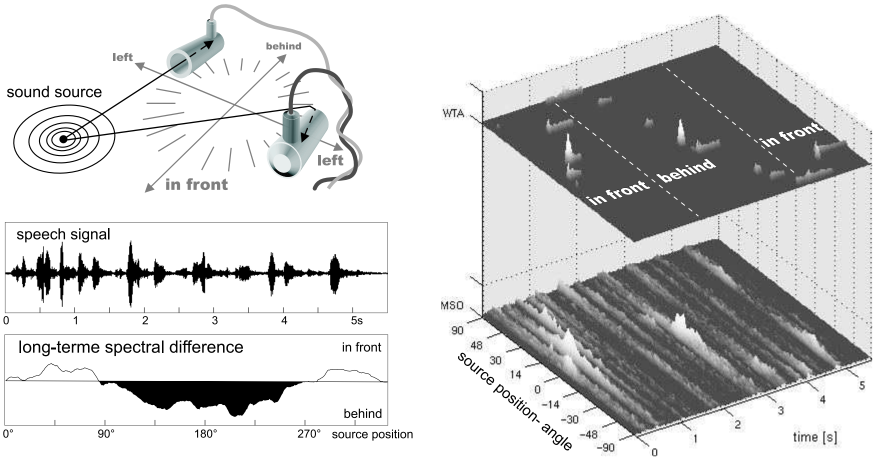

**Fig. 1. Left:** Microphone configuration and 360-degree localization of a speaker moving as described in the text. **Right:** Visualization of the IC output (bottom) and the resulting activity in the WTA (top).

## 3 Concept for a Sub-Cortical Auditory-Visual Integration

Usually, spatial perception and attention has multimodal character, whereas hearing and vision seem to be complementary strategies. 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 do not have 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 [10]. 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 multimodal model, although we also consider to integrate higher-level features as an option in concrete applications. According to [11], at least the following properties of the representation and integration of multiple sensory inputs in the SC had to be considered in the model architecture:

*Superficial SC (SCs) is responsive only to visual and especially to moving stimuli.* Counterpart of a retinotopically ordered map in SCs is a one-dimensional map of horizontally arranged intensity differences, provided by a wide-angle vision system. Presumed, that auto-motion is omitted during sensory recording, the intensity differences are coding scene motion.

*(i) Deep layers of the SC (SCd) respond to visual, somatosensory, auditory and multimodal 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 might 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.

*Maximal enhancement occurs with minimally effective stimuli.* Especially for a strong bimodal or unimodal activation, the WTA response is limited by global inhibition and by the sigmoidal output of the neurons. With a suitable set of network parameters, the combination of weak stimuli should show a greater enhancement than adding a second activation to a WTA process, that is already *saturated* by one strong stimuli.

*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 [9]. To achieve map alignment every eye-specific motor-command must cause an adjustment in the auditory map. In the model, this is realized by a controlled change of the weights of the ICx–SCd projection (fig. 2).

Similar to [7], we use a modification of the network-model of the auditory inferior colliculus to realize the multimodal map. A nonlinear notation can be given as a bimodal version of the standard dynamic field of Amari-type [1]:

$$\begin{aligned} \tau \frac{d}{dt} z(r,t) = {} & -z(r,t) + c_A x_A(r,t) + c_V x_V(r,t) \\ & -c_i \int y(z(r,t))dr + c_n \int w(r-r')y(z(r',t))dr' \end{aligned}$$

The state $z(r,t)$ of a neuron at position $r$ is depending on three components: the weighted bimodal inputs $x_A$ and $x_V$, global inhibition according to the integrated network output and lateral feedback from neighboring positions $r'$. All neurons have sigmoidal output, calculated by the Fermi-function: $y(z(r,t)) = (1+exp(-\sigma \cdot z(r,t)))^{-1}$.

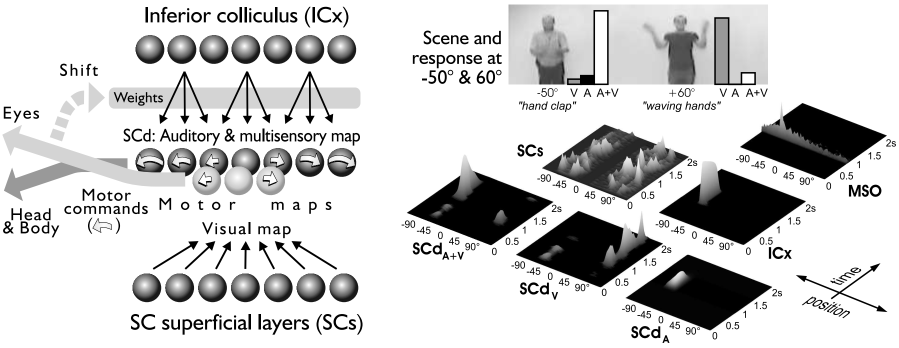

**Fig. 2. Left:** A simplified and universal model of the superior colliculus, which satisfies the properties, mentioned in the text. Further simplifications can be made, if no separate camera turns are possible (no separate SCd visual map and static ICx-SCd-projections) or if an omnidirectional camera is used (modified SCs-SCd projection). **Right:** The representation of exemplary bimodal (hand clap) and simultaneous unimodal stimuli (waving hands) is shown for the auditory, visual and multimodal maps (MSO denotes the binaural cross-correlator). To evaluate response enhancement, the SCd-map is computed three times: with only visual (V), only auditory (A) and bimodal input (A+V).

# 4 Experiments and benchmarks

## 4.1 Database concept for flexible audio-visual superposition

Since we are not processing highly specific features but modelling early stage mechanisms, the effects and properties of our simulations have to be basic and universally valid. The question, how to prove the general validity of the model in diverse and real situations, is a critical point, not only in this study. A common technique is to use recordings or realtime experiments in a restricted environment (lights and noise-bursts in a dark, anechoic chamber) for a detailed analysis. However, the relevance of those analysis to real situations must be doubt for several reasons: In reality, the stimuli are not point-like and, simply depending on the situation and the distance to the observer, occur in very different shapes, characteristics and dynamics. Another practice, which is widely used in the fields of neuroinformatics and artificial intelligence, is to generate simulations of the environment. Virtual experiments can be repeated and varied easily, but have the drawback of providing less complexity than real sensory inputs.

To overcome this dilemma, we combine recordings of real situations and off-line simulations in a novel approach. Separate recordings of sounds and visual scenes are stored in a database and can be assembled to randomly arranged but reasonable situations. The type of stimuli and scenes targets typical situations in man-machine-interaction, while the database can be extended for another purpose or to reproduce other simple multimodal setups (e.g. such as in [7]). Up to now, the visual stimuli include local motion (gestures, single hand-claps, waving hands) and translations (people walking by, getting closer or away). Since motion coded by intensity differences is the only visual cue in the model, the experiments become widely independent from object color, illumination or background texture. For the acoustic database, we recorded different words and claps and reproduced them from a number of angles in a lecture hall ($\approx 300m^2, 1sec$ reverberating time). Yet the current database, including five persons, 14 visual and 10 acoustic events offers a combinational variety, that is suitable for reproducible statistical interpretations. To share ideas and experiences, the data are public available from `http://cortex.informatik.tu-ilmenau.de/~ schauer` , where also a more detailed description of the experimental setup is provided.

### 4.2 Results

The novel database concept enables a new method of testing and evaluating our model, already in respect to real applications. Beside single experiments, it is possible to inquire benchmarks and statistical analysis. In the context of our simulations, we define a benchmark as a number of repetitions of a multimodal experiment, where certain temporal and/or spatial parameters of the scene vary. For every scene, the multimodal part of the model is simulated three times: with only visual, auditory and multimodal input. Based on these results, the amount of response enhancement can be measured in a similar way to neurophysiological recordings as $((CM - SM_{max}) \cdot 100)/SM_{max}$, where $CM$ is the combined-modality response and $SM_{max}$ is the response to the most effective single-modality stimulus [10]. The response itself is the temporal integration of the activity at one position in the spatial map. With the help of benchmark tests, it was possible to demonstrate major properties of the sensory representation in our biologically inspired model:

*Response enhancement* is performed for correlated multimodal stimulation. (fig. 3) The amount of response enhancement and the spatial disparities and effective time windows in which this effect occurs, are plausible compared to biological perception and can also be adjusted to the demands of specific applications. E.g. the spatial window for grouping auditory and visual stimuli is depending on the minimum object distance and can be realized by applying large receptive fields.

*Response depression* occurs for simultaneous but spatially separated events. In the model, this property is based on the global inhibitory mechanism of the WTA-network. In a corresponding benchmark (not plotted) the depression ranges from 10-60%.

*Maximal enhancement occurs with minimally effective stimuli.* It was shown, that a typical WTA-process is suitable to cause an inverse proportionality of single modality effectiveness and multimodal response enhancement. (fig. 4)

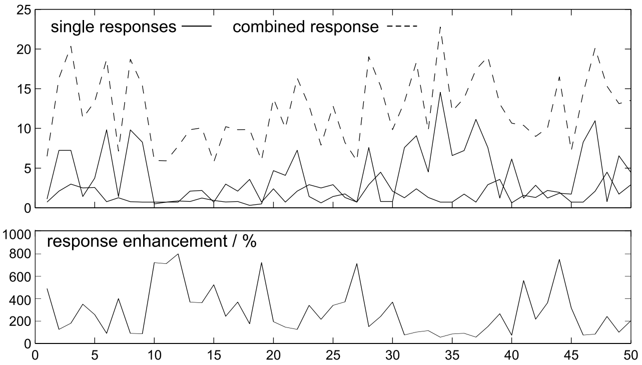

**Fig. 3.** Benchmark of 50 different multimodal events (clapping/waving hands and simultaneous speech as described in 4.1). The activity at the center-position of the auditory stimulus is shown. Response enhancement was observed in all scenes and ranges from 70-800%.

## 5 Conclusions

Based on a robust sound localization and scene motion, simply coded by temporal intensity differences, it was possible to demonstrate essential properties of sub-cortical sensory integration by simulating a dynamic neural field of Amari type with bimodal inputs. Response enhancement, response depression and the relation of maximum enhancement to least effective single modality stimuli where shown.

Further, the proposed concept of generating virtual audio-visual experiments with the help of a database of real-world scenes enables a wide range of new analysis. The spatial and temporal parameters of early multisensory integration (time-windows, receptive fields) can now be discussed by means of statistical benchmarks, already in the context of real situations and concrete applications. Along with the first simulations, we gained new insights in the mechanisms of early saliency or attention. An example is the interpretation of the very large receptive fields in the SCd, that are necessary for grouping even spatially separated stimuli, if multimodal events occur close to the observer. If, e.g. command words and gestures are processed, a person's head and hands are represented in noticeable different directions, but contributing to one multimodal event. In this situation, a high spatial resolution in the auditory and visual representations would be counterproductive. In general one can assume, that the reliability of a multimodal activation is much more important than a high localization precission (as observed e.g. during saccade generation in the superficial SC).

Although it is imaginable to perform also the initiation of motor commands and map shifting on the base of the databank, we strive for a realtime capable implementation on an experimental robotic platform and tests in real man-machine communication. The practical aspect of the application of the model is an expected significant advantage in the detection and tracking of users interacting with the mobile robot.

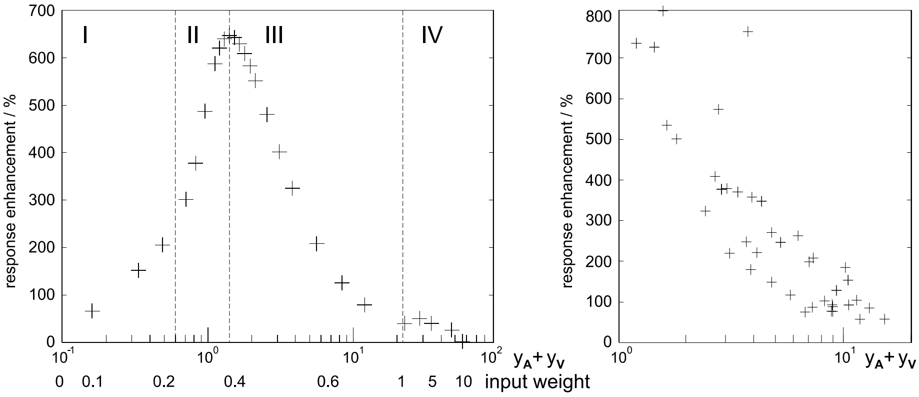

**Fig. 4.** Ineffective single-modality stimuli should produce the strongest response enhancement. **Left:** An experiment with correlated stimuli was repeated 26 times with varying input-weights of the bimodal WTA-network. The single-modality responses $y_A + y_V$ are plotted against the amount of response enhancement in the corresponding multimodal experiment. The WTA-topology and the sigmoidal output of the Amari-neurons causing four characteristic intervals: I) subthreshold range, additive combination of very weak responses, II) beginning of WTA-type behavior, III) normal WTA-process, saturation for multimodal stimulation only, IV) saturation, even for single-modality activation. **Right:** The neurophysiologic plausible condition, where single-modality activation and response enhancement are inversely proportional (interval III) was realized in the benchmark (compare fig.3)

This work was supported by Deutsche Forschungsgemeinschaft, grant to GK 164,1-96.

## References

1. Amari, S. *Dynamics of pattern formation in lateral inhibition type neural fields.* Biological Cybernetics 27:77-87, 1977.
2. Jens Blauert. *Spatial Hearing..* MIT Press, 1996.
3. G. Ehret and R.Romand The Central Auditory System. Oxford University Press, 1997.
4. Gross, H.-M., Boehme, H.-J. PERSES - a Vision-based Interactive Mobile Shopping Assistant. in: Proc. IEEE SMC 2000, pp. 80-85.
5. L.A. Jeffress. A place theory of sound localization. *Journal of Comperative Physiological Psychology*, 41:35–39, 1948.
6. John Lazzaro and Carver Mead. A silicon model of auditory localization. Neural Computation, 1(1):41–70, 1989.
7. Rucci, M., Wray, J., Edelman, G.M. Robust localization of auditory and visual targets in a robotic barn owl. Robotics and Autonomous Systems 30, 181-193, 2000.
8. 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.
9. Schiller, P.H. 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
10. Stein, B.E. and Meredith, M.A. The Merging of the Senses. The MIT Press, 1993.
11. 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