in Proc. Int. Joint Conf. on Neural Networks (IJCNN) 2004, Budapest, pp. 2523-2528

Design and Optimization of Amari Neural Fields for Early Auditory–Visual Integration

C. Schauer and H.-M. Gross
Dept. of Neuroinformatics, Ilmenau Technical University, D-98684 Ilmenau, Germany
E-mail: Schauer@informatik.tu-ilmenau.de

***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. After introducing a modified Amari-neural field architecture in the bimodal model, we place emphasis on a novel method of evaluation and parameter–optimization based on biology-inspired specifications and real-world experiments.**

I. 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, disadvantageous acoustic conditions and effects like echos and reverberation 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 advantage 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, Edelman 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.

In our approach, the biological model is not only considered for the design of neuron–models and network–topologies in the artificial system. We also use concrete neurological findings like the recording of multimodal Response Enhancement as criteria for an abstract quantitative validation and optimization of the model parameters.

II. Model Design

A. Binaural Model

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 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 the ITD processing neural circuits has been the starting point of our modeling.

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).
2) For each frequency channel, the signals 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. As the result of a first winner-take-all (WTA) process, only one direction will be dominant at a time. (Model of the external Inferior Colliculus, ICx)
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.

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 (figure 1), where the tracked speaker position is traveling once around the microphones (performed in the lecture hall without background noise).

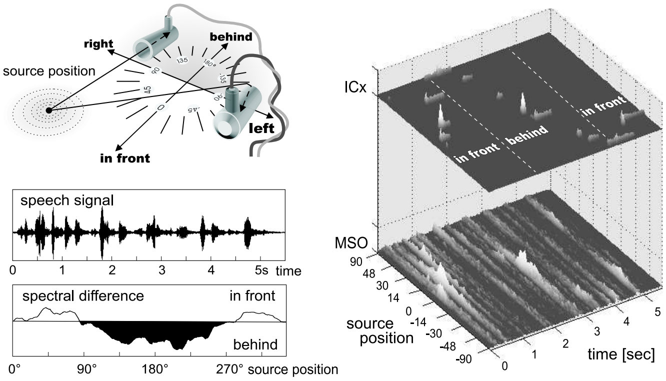


Fig. 1. **Left:** Microphone configuration and 360-degree localization of a speaker moving as described in the text. **Right:** Visualization of the MSO output (bottom) and the resulting activity in the WTA (ICx-model, top).

B. Concept for a Subcortical Auditory-Visual Integration

Usually, spatial perception and attention have 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 the frequencies of acoustic events, 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:

a) 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.

b) (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 might cover just a part of the resulting multisensory map.

c) 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.

d) Maximal enhancement occurs with minimally effective stimuli: Especially for a strong activation, the WTA response is limited by global inhibition and by the sigmoidal output of the neurons. With a suitable set of network parameters, weak bimodal inputs should show a greater enhancement than the combination of stronger stimuli, which lead to a more *saturated* WTA–process.

e) 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.

f) 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).

Like Rucci, Edelman and Wray in [7], we use similar network–types to model the auditory Inferior Colliculus and to realize the multimodal map. A nonlinear notation can be given as a dynamic field of Amari-type [1], modified by a bimodal input:

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

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: $y(z(r,t)) = (1+exp(-\sigma\cdot z(r,t)))^{-1}$.

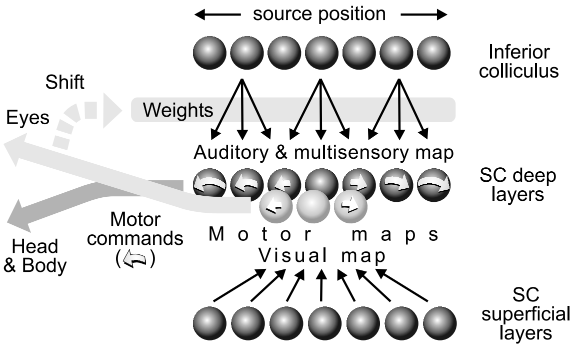


Fig. 2. A simplified and universal model of the Superior Colliculus, which satisfies the properties, mentioned in the text. Further simplifications can be made, if no separate camera turns are possible (no separate SCd visual map and static ICx-SCd-projections) or if an omnidirectional camera is used (modified SCs-SCd projection).

For every experiment, 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 (RE) can be measured in a similar way to neurophysiological recordings as:

$$RE = (CM - SM_{max})\cdot 100/SM_{max}$$

where CM is the combined-modality response (y_{A+V}) and SM_{max} is the response to the most effective single-modality stimulus ($max(y_A, y_V)$) [10]. The response itself is the temporal integration of the activity at one position in the spatial map.

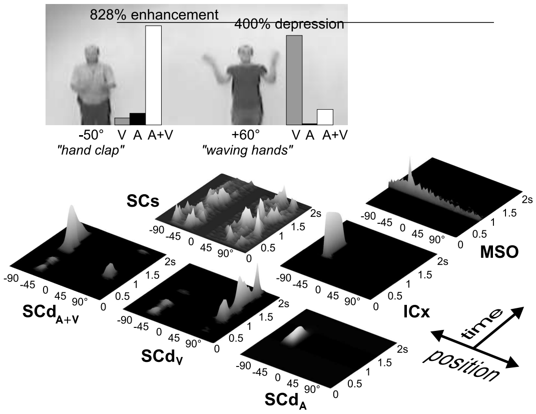


Fig. 3. The representation of exemplary bimodal (hand clap) and simultaneous unimodal stimuli (waving hands) is shown for the auditory, visual and multimodal maps (MSO denotes the binaural cross-correlator). The spatial resolution is given by the audio-recording setup (39 steps between 90 degrees left and right). To evaluate the amount of Response Enhancement, the SCd-map is computed three times: with only visual (V), only auditory (A) and bimodal input (A+V).

Below it is normalized over the duration of the experiment (usually 1-2sec.). The relation of single modality effectiveness and multimodal enhancement can be illustrated if we plot the sum of the unimodal responses $y_A + y_V$ against the amount of the Response Enhancement in the corresponding multimodal case. If we further vary the stimulus–intensity, a characteristic curve is formed for every experiment (fig.4). Based on the results of some random samples (represented by the curves in fig.4) there is evidence, that the Amari–field can perform the desired features robust and in a wide dynamic range.

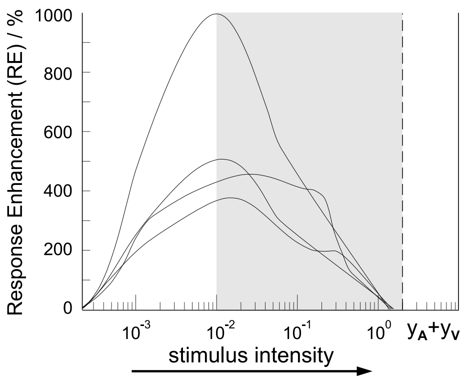


Fig. 4. Ineffective single-modality stimuli should produce the strongest Response Enhancement. Four experiments with correlated auditory and visual events were repeated with varying stimulus intensity. The single-modality responses $y_A + y_V$ are plotted against the amount of Response Enhancement in the corresponding multimodal case. The Amari–field dynamics and the sigmoidal output of the units causing characteristic intervals: After a subthreshold range with additive combination of very weak responses, the sharp rise in the output–function and the feedback in the network produce large enhancement values. If the stimuli become even stronger, saturation due to the limited output and the global inhibition is observed – first for multimodal stimulation only, and finally even for single-modality activation. The marked interval, where the RE-value seems to be generally decreasing, defines a range on the horizontal axis between 1–100% of the theoretical, saturated output of a neuron.

III. Database concept for flexible audio-visual superposition

Since we are not processing highly specific features but modeling 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 characterizes 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 people, 14 visual and 11 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.

IV. Evaluation of model behavior

The novel database concept enables a new method of testing and evaluating our model, already with respect to real–world 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. If the database covers the complexity of the potential environment, we assume that a model which performs well in the benchmark will do the same in the real application.

Now the missing link in the concept is to answer the question, what is good performance in a certain situation? For the conventional problem of object recognition an appropriate method is to distinguish between wrong and right localizations and calculate rates of successful recognitions. Here such an approach fails, because we neither are able to define objects nor decide what the right localization is in the one or the other situation. On a very early processing stage (which is modeled here), the idea of objects as well as the interpretation of stimuli–directions in the context of concrete situations are irrelevant. Since higher cognitive mechanisms seem not to be involved in the forming of primary multimodal maps in the midbrain, it simply makes no sense to use these criteria for the evaluation. Ones again, the neurobiological and physiological findings provide a solution for the problem: The response properties in the multimodal SC are the result of an evolutionary optimization – a process that was and is based on the same spatial and temporal stimuli–dynamics we have to face in our perspective application (e.g. speech and gestures). If we can quantify the multimodal properties and use them for the optimization of the SC-model, we can expect a robust model–behavior in real and natural situations.

In the following, we translate the qualitative descriptions and quantitative findings about the multimodal SC-maps into a number of optimization criteria. We start our consideration with the diagram in fig. 4. While the result of one experiment marks a point, given by $y_A + y_V$ and the corresponding RE-value, a benchmark of some hundred experiments forms a scatter plot. The noise or variation within this plot is caused by the different stimulus intensities and the more or less distinct spatial and temporal disparities in the scenes. Words are spoken with different loudness, gestures were performed rapidly or slowly, the image contrast changes with lighting, the positions of moving hands differ more or less from the head-position (which gives the auditory localization) and the time flow of gestures and sounds is a very stochastical process. In the evaluation step, we demand not less than the basic multimodal features to be reflected in the ($y_A + y_V$)-to-RE scatter plot of a benchmark. To test this hypothesis and to define boundary condition for the behavior of the model we use 5 criteria:

- **The maximum criterion:** what is the strongest enhancement, observed in all experiments of the benchmark
- **The mean criterion:** the mean of the all RE-values specifies the overall amount of the multimodal enhancement.
- **The orientation criterion:** To characterize the relation between single–modality effectiveness and multimodal enhancement, orientation and shape of the scatter–plot are crucial. A comparatively easy way to quantify these features is the Principal Component Analysis (PCA) of the scatter–plot. The Eigenvectors of the covariance matrix of the plot indicate the alignment of the data. The orientation criterion is calculated as the deviation of the strongest component from the diagonal of the normalized and centered scatter–plot (fig.5).
- **The shape criterion:** The relation of the Eigenvalues of

the covariance matrix is measuring how strong or reliable the single–/multimodal relation is. A shape-factor of 1 characterizes a *round* scatter–plot, which is not significant, larger values correspond to a dominant component in the distribution, i.e. a clear orientation of the plot.

- **The single modality criterion:** to assure that the Amari–field is performing a reasonable winner–take–all behavior, we ask for mean $y_A + y_V$-values larger than a certain minimum.

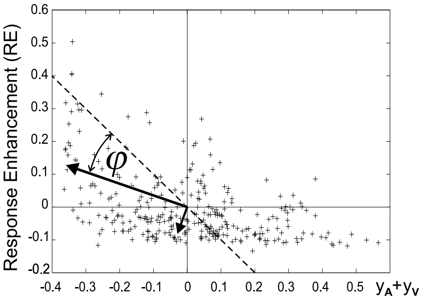


Fig. 5. The orientation and shape criteria. Normalized and centered data of a benchmark are used. The length of the arrows corresponds to the Eigenvalues of the covariance matrix of the scatter–plot, their directions are given by the Eigenvectors.

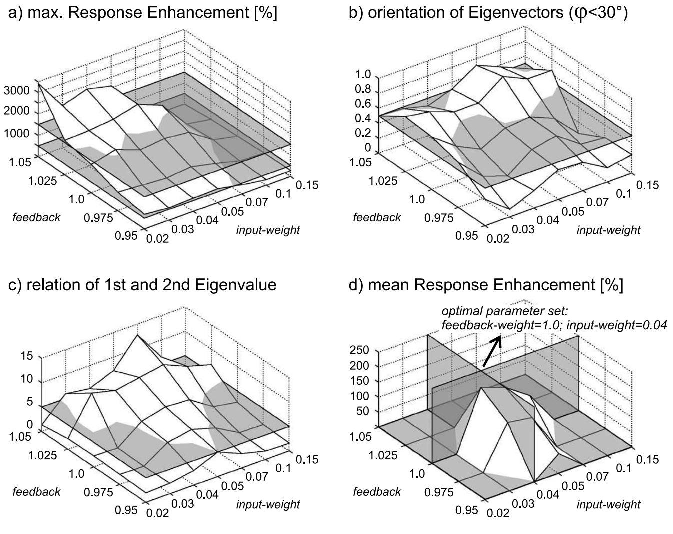
V. Amari–field optimization

The benchmarks over real audio-visual stimuli and the evaluation criteria described above can be utilized as a universal framework for the optimization of multimodal maps. Since we use a recurrent Amari-field to model the multimodal integration, the demand for reasonable or optimal topologies and parameters is of particular interest. The principle suitability of the network could be shown for random samples in fig.4, but WTA-type structures are known to be hard to analyze and adjust. We have to face variations in a high dimensional parameter space and no direct connection exist between a contrast function (or criterion) and the parameters to find an optimal operating point analytically.

Beside the evaluation criteria further constraints help to reduce the degree of freedom in the parameter space. The width of the recurrent connections of the Amari-field is controlling the way, how different input shapes are processed. The more neighboring units lay within the excitatory feedback, the broader the visual stimuli or the audio-visual disparity for a single multimodal event can be. Also the time-constant of the dynamic Amari–neurons can be fixed before the parameter search. Depending on the time-window in which multisensory interaction should occur, the time-constant was set to 250 msec in all experiments.

Other parameters like the slope of the neurons output–function or the weight on input, feedback and global inhibition are harder to adjust. In a first step, we alternated the parameters for a small number of experiments, until we found broad intervals, where the model shows acceptable WTA-behavior and the characteristic curves of fig.4 overlay as much as possible. After that, we permuted the remaining parameters (e.g. the weights for input, feedback and inhibition) in discrete gradations and run the same benchmark for each parameter set. This way a $(y_A + y_V)$-to-RE scatter–plot and the corresponding criteria are calculated on every discrete point in the remaining 3- or 4-dimensional parameter space. For a better visualization, we omit the inhibitory weight in the diagram and just plot the results for varying input- and feedback-weights and selected criteria in fig. 6. To find an optimal parameter set, we choose the mean Response Enhancement as the final optimization criterion and used the others to define a suitable subspace where a number of constraints were fulfilled: a) the highest observed enhancement values should be larger than 500% but smaller than 1500%, b) the orientation of the scatter–plot should be close to the diagonal (less than 30 degrees deviation, compare fig.5), c) the orientation should also be significant, i.e. the scatter–plot has to be narrow (factor between 1st and 2nd Eigenvalue > 5), in addition the mean single modality activation should be at least at 10% (single modality criterion, not plotted). On all points in the remaining subspace (fig. 6d), we determined the highest mean RE-value and thus got the final optimal parameter set.

The described method is universally applicable to other network types. Provided that the criteria change slowly and continuously in the parameter space, it can run automatically and is cascadable. On a common one-processor PC, a MATLAB-implementation is finding an optimum for 4 parameter (each with 10 variation) and 100 audio-visual experiments per step in about three days.


a) max. Response Enhancement [%]

b) orientation of Eigenvectors (φ<30°)

c) relation of 1st and 2nd Eigenvalue

d) mean Response Enhancement [%]

Fig. 6. Visualization of the evaluation criteria and the constraints which define a valid parameter–subspace in terms of a plausible audio-visual integration. Other Amari-field parameters were: sigmoidal output with σ=8 and inhibitory weight=0.2. The feedback vector was a normalized Gaussian with r=10.

Using the optimal parameter set, we observed not only a good winner-take-all behavior of the Amari-field but also multimodal features which correspond almost perfect to the results of known neurophysiological studies:

- **Response Enhancement** is performed for correlated multimodal stimulation (fig.7). 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. That means 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 and wide feedback connections.

- **Maximal enhancement occurs with minimal effective stimuli.** It could be shown, that a typical WTA-process is suitable to cause an inverse proportionality of single modality effectiveness and multimodal Response Enhancement. (fig.7)
- **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), depression up to 60% was observed.

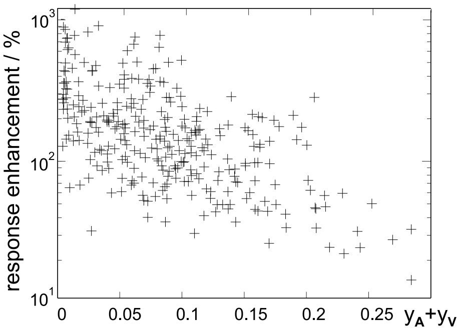


Fig. 7. Visualization of the benchmark for the optimal parameter set, including 300 experiments. The maximum multimodal enhancement was 1200%, the average value 200 %. The scatter–plot is oriented almost diagonal in the range of values and the inverse relation of single-response and multimodal enhancement is significant (factor 5.8 between 1st and 2nd Eigenvalue of the covariance matrix of the scatter–plot data).

VI. 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 methods and analysis. The spatial and temporal parameters of early multisensory integration (time-windows, receptive fields) can now be investigated by means of statistical benchmarks, already in the context of real situations and concrete applications. A key idea of this concept is to replace traditional specifications for the object recognition task by response properties, observed in the multimodal parts of the Superior Colliculus. Exemplarily we realized a successful high–dimensional parameter optimization in a hard–to–adjust recurrent network based on multiple bio-inspired criteria. In a next step, we are going to integrate a Response Depression criterion to complete the evaluation and optimization concept.

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 the same 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 precision (as observed e.g. during saccade generation in the superficial SC or in the auditory maps of ICx).

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 robot 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 a mobile robot [4].

Acknowledgment:

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

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 Comparative 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