11

Towards Computational Modelling of Neural Multimodal Integration Based on the Superior Colliculus Concept

Kiran Ravulakollu, Michael Knowles, Jindong Liu, and Stefan Wermter

University of Sunderland
Centre for Hybrid Intelligent Systems
Department of Computing, Engineering and Technology
St Peters Way, Sunderland, SR6 0DD, UK
www.his.sunderland.ac.uk

Abstract. Information processing and responding to sensory input with appropriate actions are among the most important capabilities of the brain and the brain has specific areas that deal with auditory or visual processing. The auditory information is sent first to the cochlea, then to the inferior colliculus area and then later to the auditory cortex where it is further processed so that then eyes, head or both can be turned towards an object or location in response. The visual information is processed in the retina, various subsequent nuclei and then the visual cortex before again actions will be performed. However, how is this information integrated and what is the effect of auditory and visual stimuli arriving at the same time or at different times? Which information is processed when and what are the responses for multimodal stimuli? Multimodal integration is first performed in the Superior Colliculus, located in a subcortical part of the midbrain. In this chapter we will focus on this first level of multimodal integration, outline various approaches of modelling the superior colliculus, and suggest a model of multimodal integration of visual and auditory information.

1 Introduction and Motivation

The Superior Colliculus (SC) is a small part of the human brain that is responsible for the multimodal integration of sensory information. In the deep layers of the SC integration takes place among auditory, visual and somatosensory stimuli. Very few types of neurons, such as burst, build-up and fixation neurons are responsible for this behaviour [4, 10]. By studying these neurons and their firing rates, integration can be successfully explored. The integration that takes place in the SC is an important phenomenon to study because it deals with different strengths of different stimuli arriving at different times and how the actions based on these stimuli are generated. There is evidence that when two different stimuli are received at the same time, the stronger signal influences the response accordingly based on Enhancement and Depression Criteria [3]. A better understanding of multimodal integration in the SC not

only helps in exploring the properties of the brain, but also provides indications for building novel bio-inspired computational or robotics models.

Multimodal integration allows humans and animals to perform under difficult, potentially noisy auditory or visual stimulus conditions. In the human brain, the Superior Colliculus is the first region that provides this multimodal integration [23]. The deep layers of the Superior Colliculus integrate multisensory inputs and generate directional information that can be used to identify the source of the input information [20]. The SC uses visual and auditory information for directing the eyes using saccades, that is horizontal eye movements which direct the eyes to the location of the object which generated the stimulus. Before integrating the different modalities the individual stimuli are preprocessed in separate auditory and visual pathways. Preprocessed visual and auditory stimuli can then be used to integrate the stimuli in the deep layers of the SC and eventually generate responses based on the multimodal input.

The types of saccades can be classified in different ways [39] as shown in Figure 1. Most saccades are reflexive and try to identify the point of interest in the visual field which has moved due to the previous visual frame changing to the current one [20]. If no point of interest is found in the visual field, auditory information can be used to identify a source. Saccades are primary actions which in some cases are autonomous and are carried out without conscious processing in the brain. When there is insufficient information to determine the source based on a single modality, the SC uses multimodal integration to determine the output. Saccades are the first actions taken as a result of receiving enough visual and auditory stimuli.

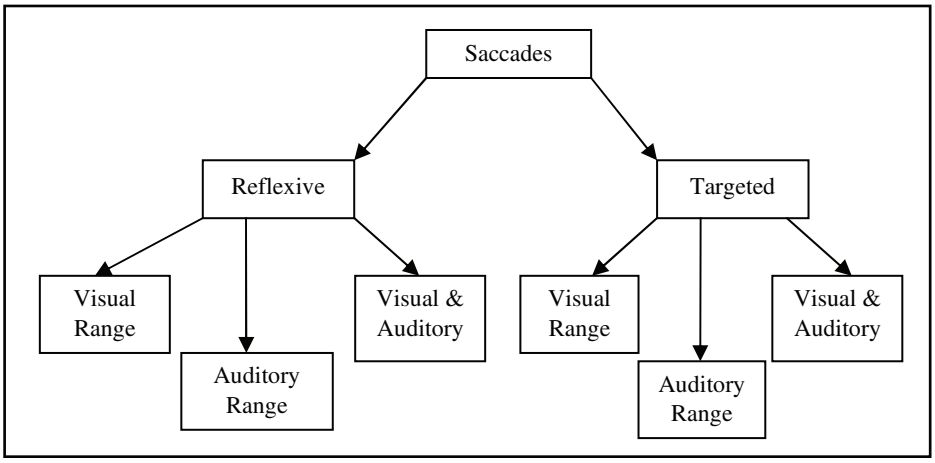


Fig. 1. The different types of saccades that are executed by the eyes of a mammal

It is known that the SC plays a significant role in the control of saccades and head movements [3, 23, 33]. In addition to the reflexive saccades, saccades can also be targeted on a particular object in the visual field. In this case, the SC receives the visual stimulus at its superficial layers which are mainly used to direct the saccades to any change in the visual field. In a complementary manner, the deep layers of the SC are capable of receiving auditory stimuli from the Inferior Colliculus and other

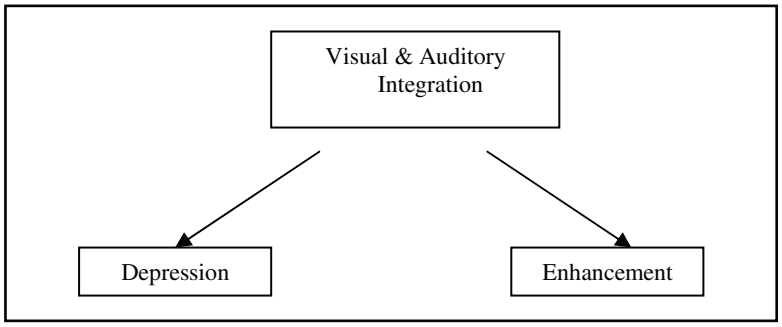


Fig. 2. Types of effects in multimodal integration after integrating the visual and auditory fields.

cortical regions. The deep layers of the Superior Colliculus are also responsible for integrating multisensory inputs for the generation of actions. The SC is capable of generating output even if the signal strength of a single modality would not be sufficient or too high to trigger an action (see Figure 2). The *enhancement* is increasing the relevance of a particular modality stimulus based on the influence of other modalities while *depression* is decreasing the relevance of a particular modality, in particular if the stimuli disagree.

Several authors have attempted to examine multimodal integration in the Superior Colliculus [2, 4, 5, 10, 27, 37, 51, 55]. Different approaches have been suggested including biological, probabilistic and computational neural network approaches. Of the different researchers that pursued a probabilistic approach, Anastasio has made the assumption that SC neurons show inverse effectiveness. In related probabilistic research, Wilhelm et al used a Gaussian distribution function along with a preprocessing level for enhancement criteria in SC modelling. Other researchers like Maren and Dominik [30] and Mass et al [17] have used a probabilistic approach in conjunction with other extraction techniques for modelling multimodal integration for other applications.

In contrast to probabilistic approaches, Trappenberg [48], Christian Quaia [4] and others have developed neural models of the SC. Christiano Cuppini [5] and Casey Pavlov [33] have also suggested a specific SC model but the model is based on many assumptions and only functions with specified input patterns. Also the enhancement and depression criteria are not comprehensively addressed. Yagnanarayana has used the concept of multimodal integration based on the superior colliculus and implemented it in various applications of an autoassociative neural network. According to J. Pavon [19], this multimodal integration concept is used to achieve efficiency in agents when co-operating and co-ordinating with various sensor data modalities. Similarly Rita Cucchiara [39] has established a biometric multimedia surveillance system that uses multimodal integration techniques for an effective surveillance system, but the multimodal usage is not able to overcome problems like the influence of noise.

In summary, while probabilistic approaches often emphasise computational efficiency, the neural approach offers a more realistic bio-inspired approach for modelling the Superior Colliculus. Therefore, in the next section our neural SC modelling methodology will be described followed by enhancement and depression modelling performed by the SC.

2 Towards a New Methodology and Architecture

2.1 Overview of the Architecture and Environment

The approach proposed in this section addresses the biological functionality of the SC in a computational model. Our approach mainly aims at generating an integrated response for auditory or visual signals. In this context, neural networks are particularly attractive for SC modelling because of their support of self organisation, feature map representations and association between the layers. Figure 3 gives an overview of our general layered architecture. Hence a two-layer neural network is considered where each layer receives inputs from different auditory and visual stimuli, and integrates these inputs in a synchronised manner to produce the output.

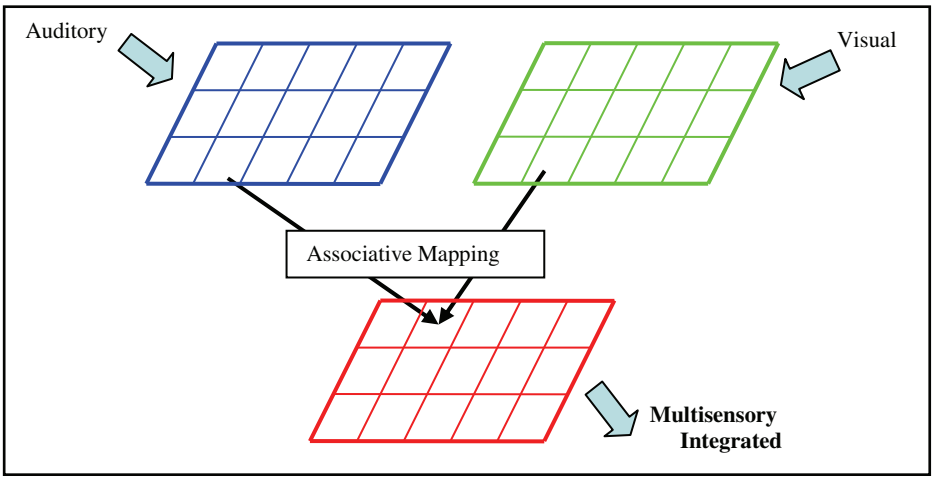


Fig. 3. A layered neural network capable of processing inputs from both input layers. Integration is carried out by an associative mapping.

For auditory input data processing, the Time Difference of Arrival (TDOA) is calculated and projected on the auditory layer. The auditory input is provided to the model in the form of audible sound signals within the range of the microphones. Similarly for visual input processing, a difference image, DImg, is calculated from the current and previous visual frames and is projected on the visual layer. Both visual and auditory input is received by the network as real time input stimuli. These stimuli are used to determine the source of the visual or auditory stimuli. Then the auditory and visual layers are associated for the generation of the integrated output. In case of the absence of one of the two inputs, the final decision on the direction of the saccade is made based only on the present input stimuli. In the case of simultaneous arrival of different sensory inputs, the model integrates both inputs and generates a common enhanced or depressed output, depending on the signal strength of the inputs. Our particular focus is on studying the appropriate depression and enhancement processes.

For evaluating our methodology we base our evaluation on the behavioural framework of Stein and Meredith [3]. Stein and Meredith's experiments are carried out examining a cat's superior colliculus using a neuro-biological/behavioural methodology to test hypotheses in neuroscience. This experimental setup provides a good starting point for carrying out our series of experiments. Our environment includes a series of auditory sources like speakers and visual sources like LEDs arranged as a semicircular environment so that each source will have the same distance from the centre covering 180 degrees of visual and auditory ranges.

For the model demonstration, Stein and Meredith's behavioural setup is modified by replacing the cat with a robot head as shown in Figure 4. As a result of visual or auditory input the robot's head is turned towards the direction of the source. The advantages of using this environment are that environmental noise can be considered during detecting and tracking and the enhancement and depression phenomena can be studied.

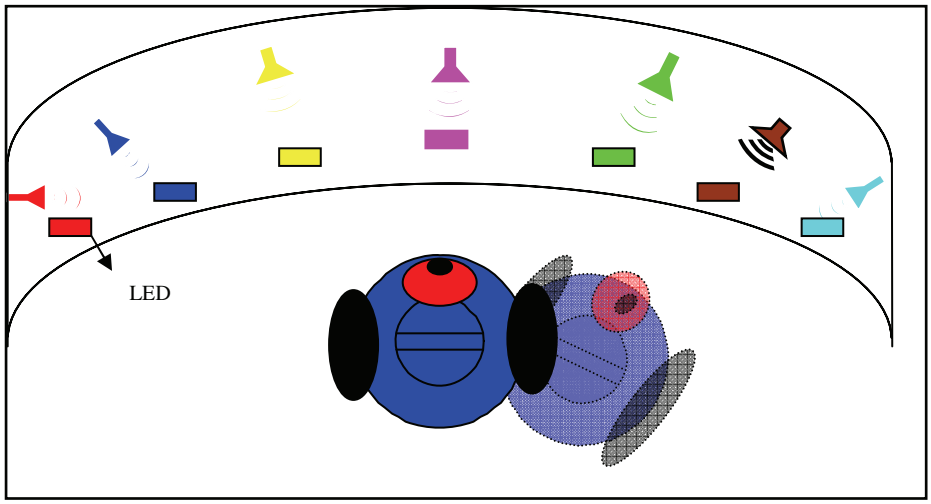


Fig. 4. Experimental setup and robot platform for testing the multimodal integration model for an agent based on visual and auditory stimuli

2.2 Auditory Data Processing

The first set of experiments is based on collecting the auditory data from a tracking system within our experimental platform setup. We can determine the auditory sound source using the interaural time difference for two microphones [31] and the TDOA (Time Difference Of Arrival), which is used for calculating the distance from the sound source to the agent. The signal overlap of the left and right stimuli allows determining the time difference.

$$TDOA = \left\{ \frac{length(xcorr(L,R)) + 1}{2} - \max(xcorr(L,R)) \right\} \times S_r$$

In the above equation '*xcorr ()*' is the function that determines the cross-correlation of the left 'L' and right 'R' channels of a sound signal. S_r stands for sample rate of the sound card used by the agent. Once the time difference of arrival is determined the distance of the sound source from the agent is calculated using the following:

$$Distance = TDOA \text{ x } sound\ frequency$$

The result is a vector map for further processing to generate the multimodal integrated output. However, for unimodal data sets, we can determine the direction of the sound source in a simplified manner using geometry and the data available including speed of sound and distance between the ears of the agent that is shown in the circular diagram below.

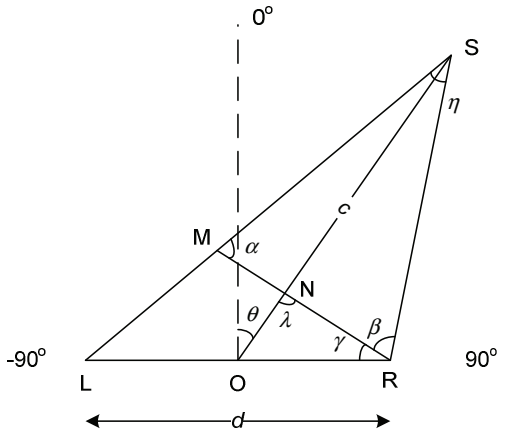


Fig. 5. Determination of sound source directions using TDOA where c is the distance of the source and θ is the angle to be determined based on TDOA and the speed of sound. The distance d between two microphones, L and R, is known.

Assume the sound source is in far-field, i.e. c>>d, then $\alpha=\beta=\lambda\approx 90$ degree and the sound source direction can be calculated using triangulation as follows:

$$\theta=\gamma=\arcsin(ML/d)=\arcsin(TDOA*Vs/d)$$

where Vs is the sound speed in air.

From the above methodology the unimodal data for the auditory input is collected and the auditory stimuli can be made available for the integration model.

2.3 Visual Data Processing

We now consider visual data processing where simple activated LEDs represent the location of the visual stimulus in the environment. A change in the environment is

determined based on the difference between subsequent images. By using a difference function, the difference between the two successive frames is calculated.

$$DImg = abs\_image\_difference(image(i), image(i-1))$$

Once the difference images are obtained containing only the variations of the light intensity, they are transformed into a vector. Once the vectors are extracted they can be used as direct inputs to the integrated neural model. However, in the case of unimodal data, difference images (DImg) are processed directly to identify the area of interest. The intensity of the light is also considered to identity an LED with larger brightness. Using this method a series of images is collected. From this vector the maximum color intensity location (maxindex) is identified and extracted. Using this information and the distance between the centre of the eyes to the visual aid, it is easy to determine the direction of the visual intensity changes in the environment using the following formula:

$$\theta = \tan^{-1}\left\{\frac{(\text{maxindex} - \text{half\_visual\_width}) \times \text{visual\_range}}{\text{visual\_width x distance\_of\_sensor\_to\_source}}\right\}$$

After running this experiment based on a set of 5 LEDs the different difference images were collected. By transforming the image on a 180 degree horizontal map with 5 degree intervals, the angle of the source is identified. For the data collected in the visual environment, 85 – 95 % accuracy is achieved for single stimulus inputs, and 79 – 85% accuracy is observed for multiple stimuli.

Later during the integration, the signal strength is also included in the network for generating the output. Stein and Meredith have previously identified two phenomena,

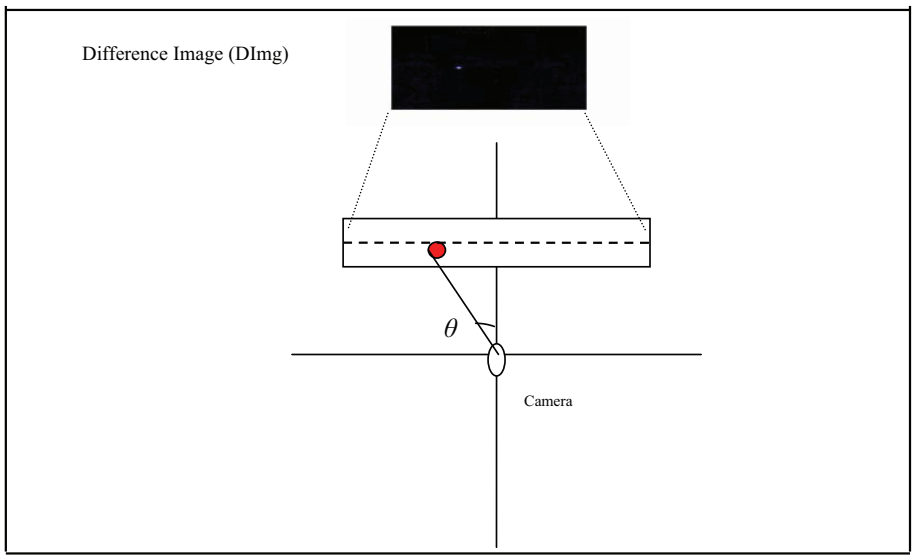


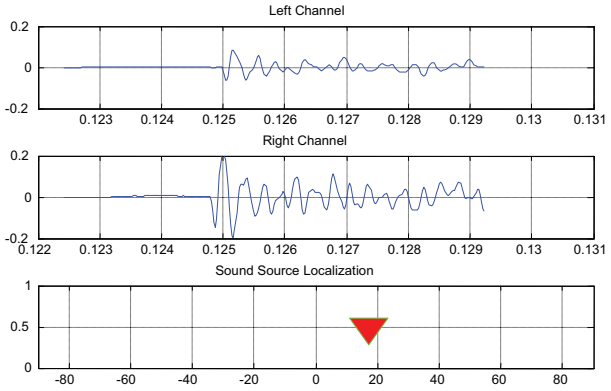
Fig. 6. Difference Image (DImg) used for scaling and to determine the location of the highlighted area of the image in which the dash line represents the length of the visual range and distance between the two cameras

depression and enhancement, as crucial for multimodal integration. In our approach we have also considered the visual constraints from the consecutive frames for confirming whether a signal of low strength is a noise signal or not. By reducing the auditory frequency to 100Hz for a weak auditory signal and by also reducing the LEDs in the behavioural environment we are able to generate a weak stimulus to study the enhancement response.

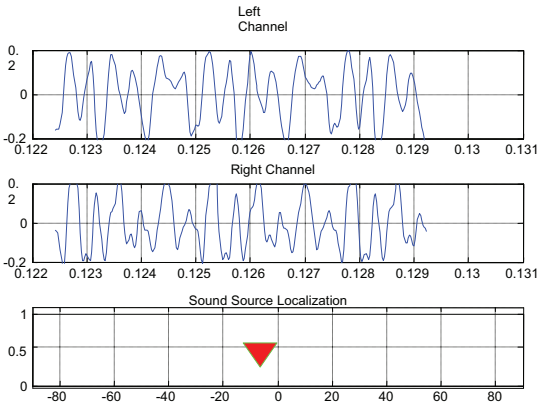
3 Simplified Modelling of the Superior Colliculus

3.1 Unimodal Experiments

First, unimodal data from auditory and visual stimuli are collected and the desired result is estimated to serve as comparison data for the integrated model. The agent is a



(a)



(b)

Fig. 7. Graphical representation of auditory localization. The behavioural environment when there are two auditory inputs. The signal received at the left and right microphone are plotted on a graph with time on the x-axis and amplitude on the y-axis. Once the TDOA is calculated and the direction of the auditory source is located it can be shown in the range from -90 to 90 degrees which in this case is identified as (a) 18 degree and (b) -10 degrees.

PeopleBot robot with a set of two microphones for auditory input and a camera for visual input.

3.1.1 Auditory Experiments

The agent used for auditory data collection has two microphones attached on either side of the head resembling ears. The speakers are stimulated using an external amplifier to generate sound signals of strength within human audible limits. For these experiments the signal levels of smaller frequencies are considered, since with these lower frequencies the multimodal behaviour can be identified and the behaviour can be studied at a more critical precision. Hence the frequencies with a range of 100 Hz - 2K Hz are used. Sound stimuli are generated randomly from any of the different speakers, and by implementing the interaural time difference method from above, the direction of the stimulus is determined. The following table provide example results of these auditory experiments of various directions and stimuli levels.

By running the above experiment for lower frequency ranges of 100 to 500 Hz with the amplitude level at 8 (as the recognition is effective at this level), initial experiments were carried out and the results are presented below. For each frequency from 100Hz, the sound stimuli are activated at angles varying from -90 to 90 degrees. Below in table 1 we show the angles computed by the tracking system discussed in the above section.

Table 1. This table depicts the accuracy level of the various frequencies from 100Hz to 500Hz.

	Actual Angle												
Freq. vs Angle	-90	-60	-45	-30	-20	-10	0	10	20	30	45	60	90
100	-81.07	-61.39	-6.29	-31.44	-21.41	-8.4	0	10.52	21.41	33.97	41.07	63.39	50.04
200	-71.07	-63.39	-42.11	-33.97	-25.97	-14.8	0	10.52	21.41	35.59	42.11	63.69	80.2
300	-76.88	-63.39	-41.07	-29.97	-25.97	-14.8	-2.09	12.4	21.41	31.44	38.95	63.39	80.00
400	-73.19	-63.39	-41.07	-75.6	-33.41	-10.52	-2.09	10.42	16.98	36.59	41.07	63.39	73.41
500	-43.9	-63.4	-17	-22.14	-17	-10.5	0	10.52	21.41	29.29	48.4	63.39	53.41

3.1.2 Visual Experiments

The camera is directed to cover 120 degrees, from -60 to 60. The series of frames collected as input from the camera are processed and the output should determine which of the LEDs is active. The frames that are captured from the camera are used to produce the difference image (DImg) which contains only the changes that have occurred in the visual environment of the agent. These difference images are calculated by measuring the RGB intensity variations from pixel to pixel in the two successive images.

Below we show how these difference images are generated based on two successive frames where the first frame is a null image with no activation and the second frame has activation at two different locations. The third image is the difference

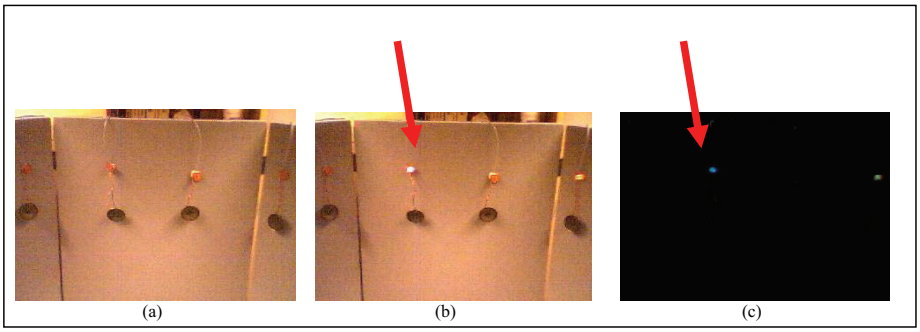


Fig. 8. (a) The visual environment without visual LED stimulus (b) The environment with visual LED stimulus. (c) The difference image (DImg).

image generated from the first two frames. This series of difference images is used to identify change in the environment. The difference images can be plotted on a plane covering -90 to 90 degrees on a special scale to intensity values of the difference image signal strength. The following image represents one such plotting of the difference image signal intensity.

Using this plot, the direction of the source from the centre of the agent is determined. The difference images are mapped on to a standard Horizontal Scale Frame (HSFr), to determine the location of the activation. The HSFr is a scale that divides the 180 degrees with 5 degrees of freedom. This scale image is different for different images.

In this horizontal scale frame the horizontal axis is divided into 10 degree intervals. Hence all the visual information that arrives at the camera of the agent is transformed into a difference image intensity plot and finally plotted on an HSFr to locate the source in the visual environment. Within this process, different auditory and visual inputs are collected and later used as a test set for the neural network that can generate multimodal integrated output for the similar auditory and visual inputs.

A synchronous timer is used to verify and confirm whether the visual and auditory stimuli are synchronized in terms of time of arrival (TOA). If the arrival of the stimuli is asynchronous then an integration of the inputs is not necessary, as the location of the source can be determined depending on the unimodal processing. In cases of multiple signals with a synchronous TOA, the *signal strength* is considered for both signals. Once the strongest signal is identified then the preference is given first to this signal and only later an additional preference may be associated with the other signal. This case occurs mainly with unimodal data such as a visual environment with two different visual stimuli or an auditory field with two different stimuli.

3.2 Multimodal Experiments

Now we focus on the combination of both auditory and visual processing as shown in Figure 4. The received auditory and visual inputs are preprocessed considering the functionality of the optic chiasm and the information flow in the optic tract. This

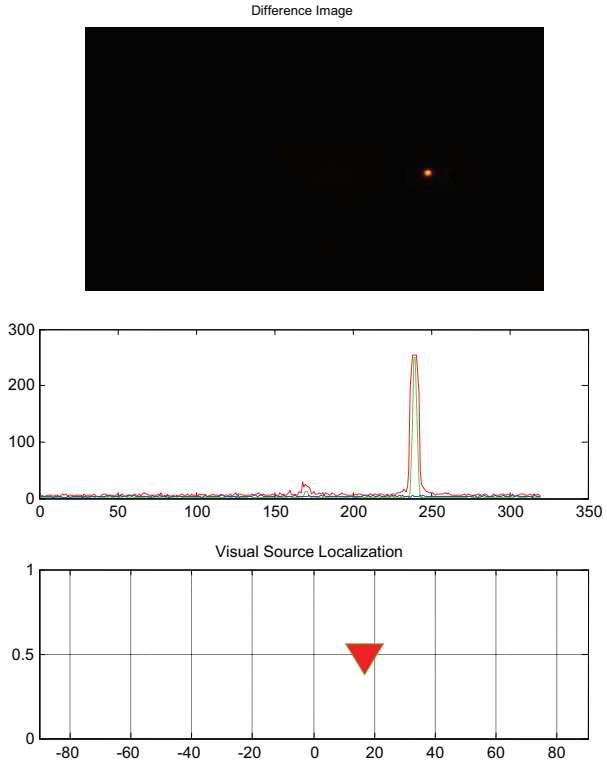


Fig. 9. (a) The difference image (DImg) is shown scaled to a unique size for all the images to standardize as a vector for the map alignment in the multimodal phase. (b) Horizontal Scale Frame image (HSFr) is a frame scale which is scaled to -90 to 90 degrees used as a reference to check in which block the enlightened part of the difference image falls.

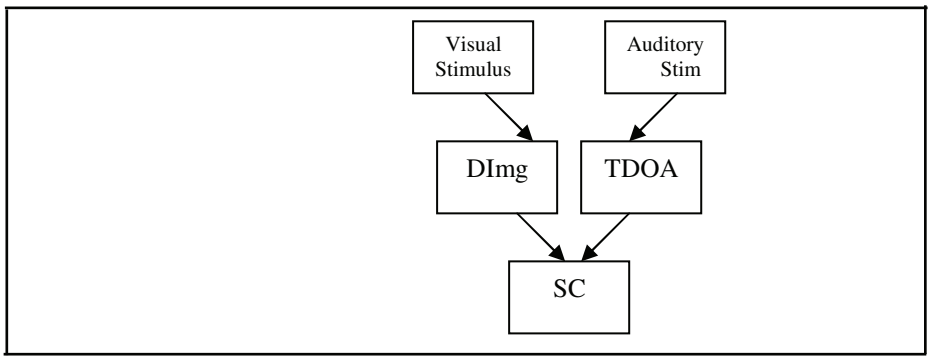


Fig. 10. Schematic representation of stimuli flow from the external environment to the superior colliculus model

preprocessing generates the difference image (DImg) from the captured visual input and the time difference of arrival (TDOA) for the auditory stimulus as shown in Figure 10. The preprocessed information enters the network which performs multimodal integration of the available stimulus and a corresponding output is generated.

The network model is mainly focused on map alignment to resemble the actual integration in the deep layers of the Superior Colliculus. In this network two types of inputs are considered: the TDOA auditory map and the DImg visual stimulus map are used for the integration. The TDOA is converted into a vector of specific size which in this case is 10 by 320 containing the information of the waveform with the intensities of the input signal. Similarly the difference image is converted into a vector of similar size containing only the information of horizontal intensity variations in the difference image.

These two vector representations are the input for the multimodal integration model. Since we are focussing on horizontal saccades, only the variations at a horizontal scale are considered at this point although this can be extended to vertical saccades. To identify the highest of multiple stimuli, a Bayesian probability-based approach is used to determine the signal strength as input to the network. A synchronous timer is used for counting the time that lapses between the arrival of the various stimuli of the corresponding senses.

The integration model is a two-layer neural network with the size 10 by 32 based on the size of the input images generated. Hence for the integration in the network the weighted average is considered as follows:

$$Integrated\ Output = ((W_V * V_I) + (W_A * A_I))/(W_V + W_A)$$

where W_v = *Visual Vector Weight,* W_A = *Auditory Vector Weight,* V_I = *normalised Visual Vector* and A_I = normalised *Auditory Vector*.

This weight function determines the weighted location of the source and provides a degree value where the source is. The difference between the two will allow the model to determine the stronger source.

Integration Case Studies:

- **Multiple visual input stimuli:** In case of more than one visual input in the environment, the difference image is generated highlighting the areas of visual interference. From the difference image the intensity of the signal is identified in terms of RGB values as shown below in figure 11. Examining the first and last spike shows that the green spike is low in intensity compared to the second one. Considering the second, the green and red spikes are high in the intensity when compared to the rest. However, the plot of the maximum values of the available RGB intensities determines the position of the source. By plotting the position onto a [-90, 90] scale the location of the source is determined, which in this example case in figure 11 is -30◦.

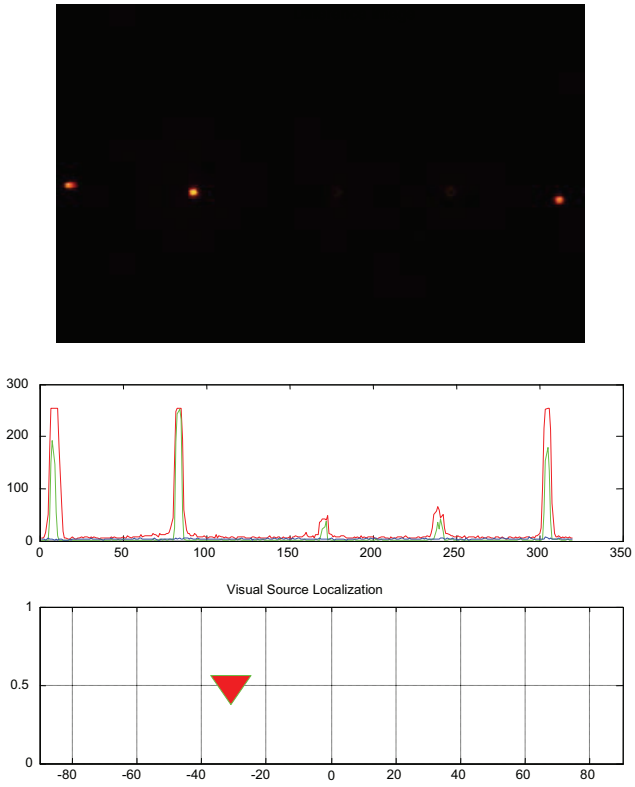


Fig. 11. The figure depicts multiple visual input stimuli received by the agent and how the visual localization is determined using the difference image and intensity plotting. The maximum intensity is identified in the second peak where all RGB values are highest when compared with the rest of the peaks.

If there are even more visual stimuli, the behaviour of our model is similar, and it calculates the intensities of the signals and plots the maximum of them in the standard plotting area [-90, 90]. In this case a close inspection of the spikes reveals the small difference that is present in the green spikes of each stimulus spike.

- **Low auditory and strong visual stimuli:** If a low intensity auditory signal and a visual signal with strong intensity value are received at the same time as input to the multimodal integration system, after verifying the time frame to confirm the arrival of the signals, both inputs are considered. After preprocessing of the signals, the signal maps are generated. In the graphs we can observe that the signal in the auditory plot has a very low intensity and the angle is determined. For the visual stimulus, the single spikes in red and green are considered for the maximum signal value. When plotted on the standard space scale, the source locations are identified as two different locations but the overall integrated location is identified as being close to the stronger visual stimulus.

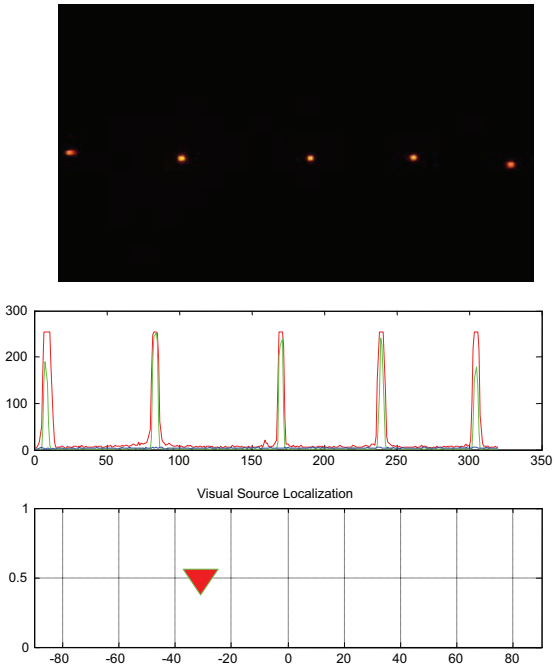


Fig. 12. The figure depicts multiple visual stimuli and how the model deals with such input. The plot shows the maximum intensity peak location and how the preferred location is determined.

- **Strong Auditory and Low Visual Stimuli:** In this case the intensity of the auditory signal is strong and the intensity of visual signal is low. For the visual stimulus, the strength of the green spikes is similar for both activation cases, while the red spikes vary. Determining the location of the two inputs individually, the locations are on different sides of the centre. When the multimodal output is generated, the location of the integrated output is close towards the auditory stronger stimulus.

The above two representative cases are observed during multimodal integration with one of the signals being very strong in its intensity. The Superior Colliculus model focuses on the stimulus with the highest intensity and therefore the integrated decision is influenced by one of the stimuli.

- **Strong visual and strong auditory stimuli:** In this case scenario, when the signals are received by the sensors, the signal intensities are calculated and the modalities are plotted on an intensity graph to determine the signal intensity. In the intensity graphs shown, the sources are located at either side of the centre and the activations are of high intensities. When the output is computed, the source is located close to the visually detected peak since the visual stimulus has greater priority than the auditory stimulus.

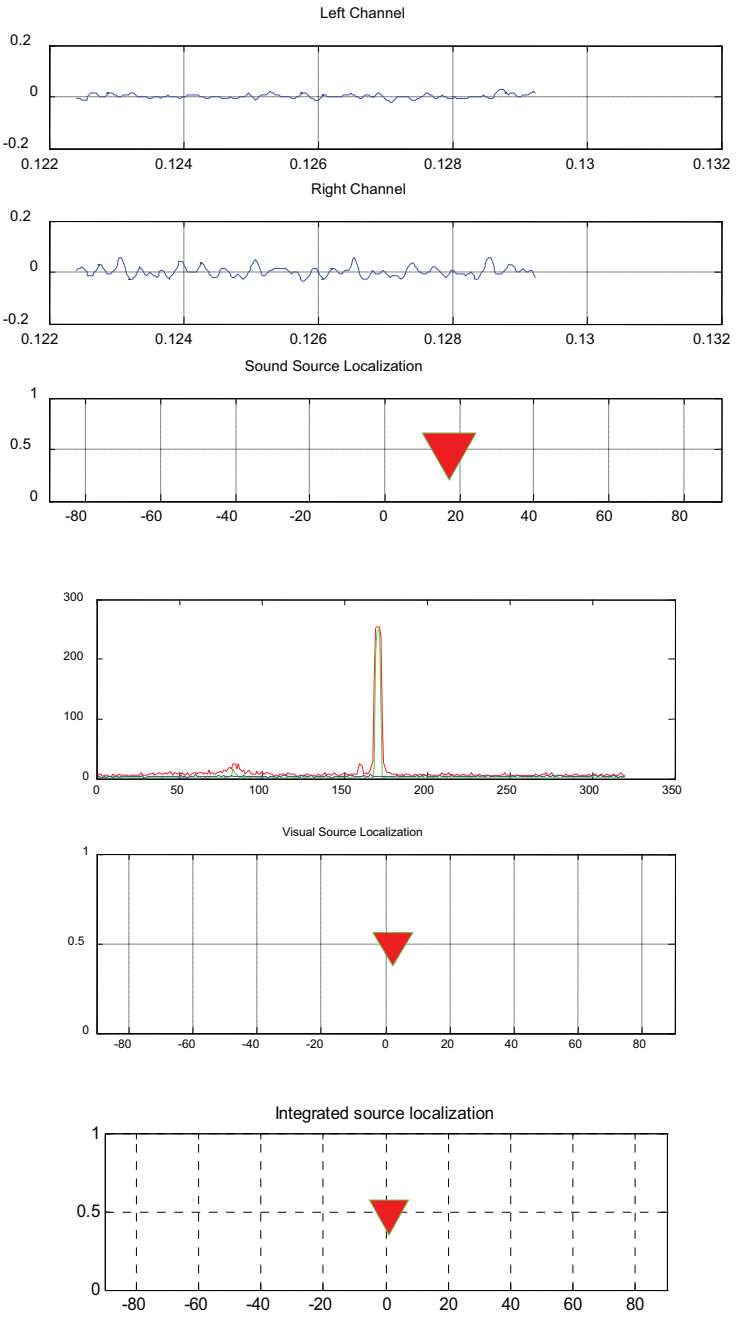


Fig. 13. Auditory and visual input with strong visual stimulus determining the main preference for the localization.

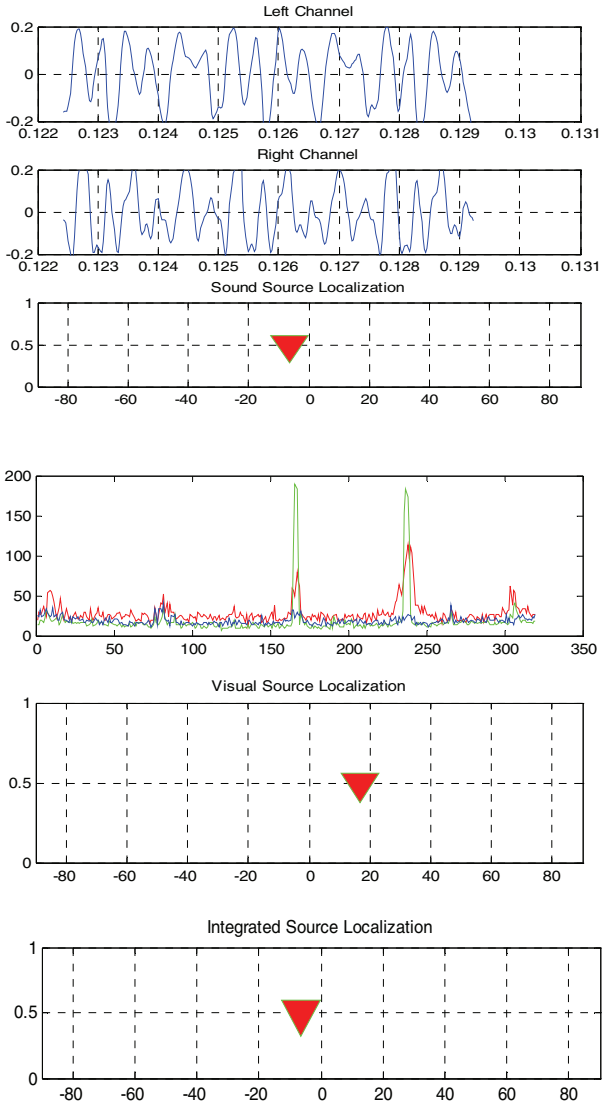


Fig. 14. Multimodal input case with strong auditory stimulus. The integrated output is biased by the intensity of the stronger stimulus which in this case is auditory.

It is not clear whether the superior colliculus will prioritize in every case, but in the case of multiple strong intensity stimuli the visual stimulus will have the higher priority while the strong auditory stimulus will have some influence on the multimodal integrated output.

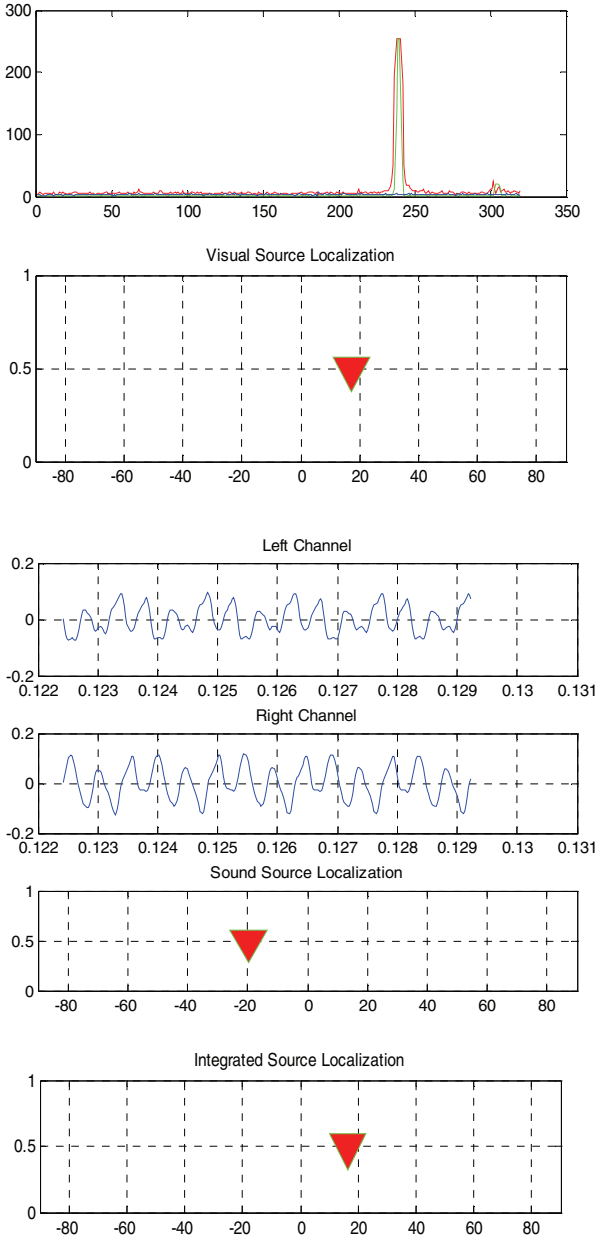


Fig. 15. Multimodal enhancement response: The integrated output is generated based on a distance function between the auditory and visual intensity.

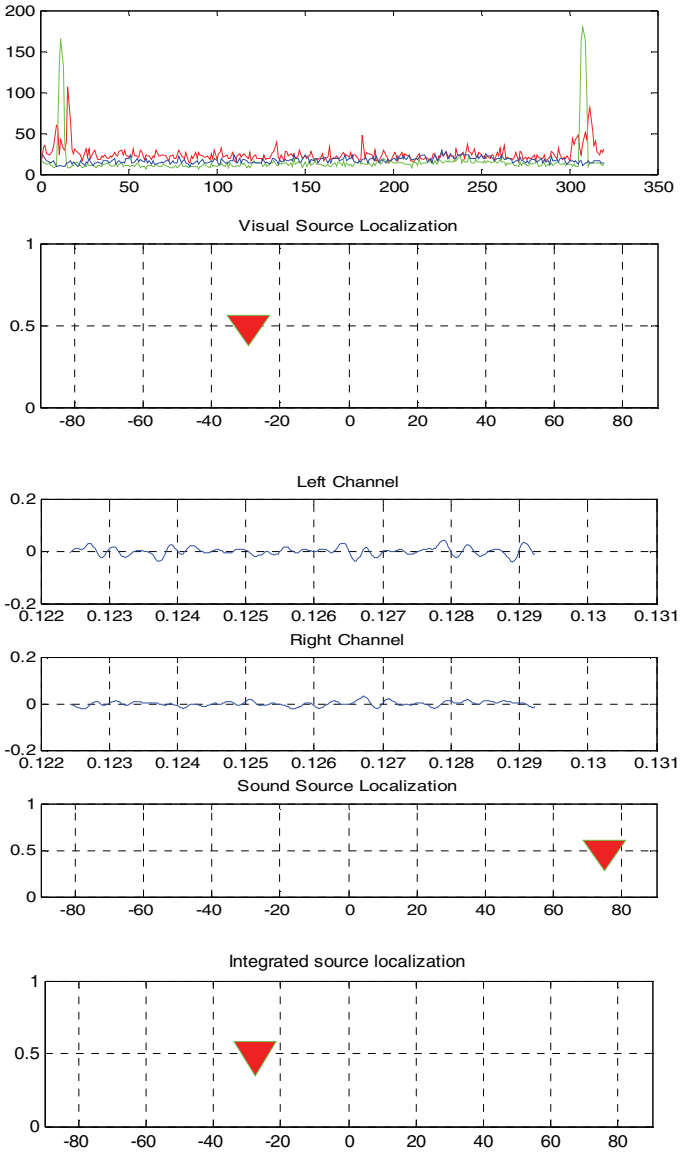


Fig. 16. Multimodal depression response: The integrated output is generated based on a distance function between the auditory and visual intensity where the auditory signal is suppressed due to its low intensity and the visual stimulus is the only input available. Hence the distance function is biased towards the visual stimulus.

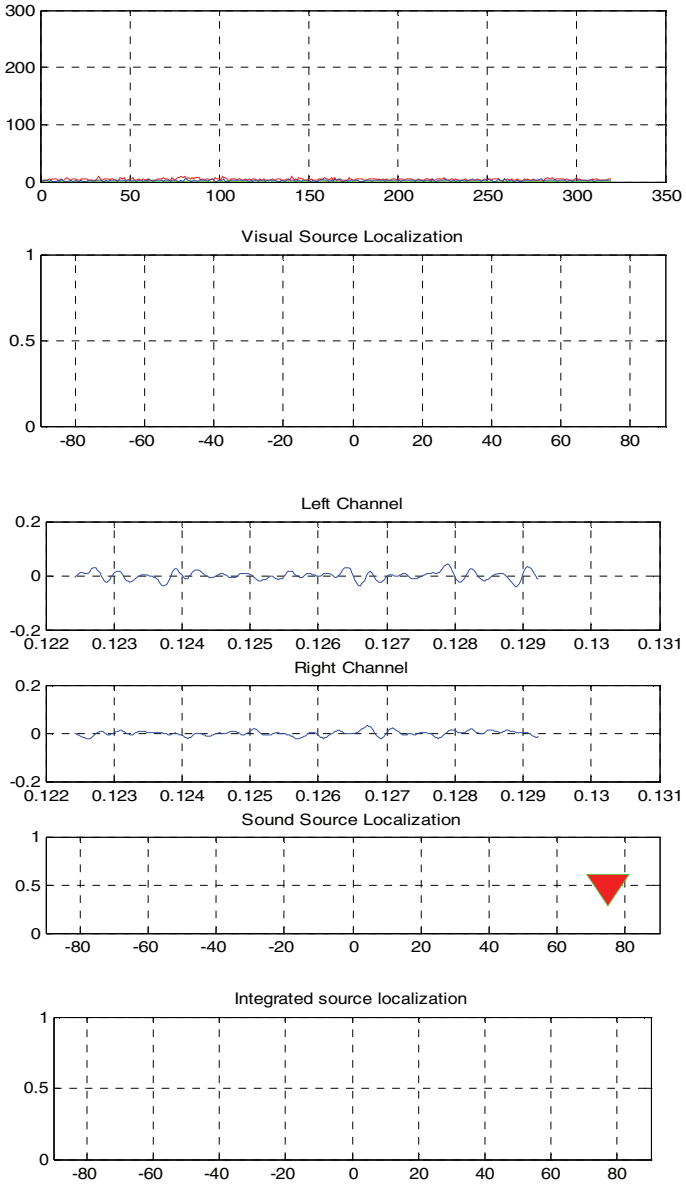


Fig. 17. Multimodal Depression Responses: a weak or low intensity auditory signal has suppressed the total multimodal response and generated a new signal that can achieve the response accurately but with a very low signal strength. This phenomenon is observed once in twenty responses, as in the remaining cases, the model tries to classify the stimuli to generate the output.

- **Low visual and low auditory:** In circumstances where both visual and auditory signals are of low intensities, the behaviour of the superior colliculus is often difficult to predict and determine. In this case, the superior colliculus can be thought of as a kind of search engine that keeps traversing the environment to locate any variations in the environment. When both auditory and visual signals are of low intensity, the SC suppresses the auditory signal due to its low intensity. Though the visual signal is low in its intensity, as far as the SC is concerned, it is the only sensory data available. Therefore, the source is identified much closer to the visual stimulus as shown in figure 16 below.

In some cases, if the stimulus is not in the range of [-30, 30] degrees, depression may occur if the visual stimulus is getting out of range or the auditory stimulus is getting less intense.

4 Summary and Conclusion

We have discussed a model of the SC in the context of evaluating the state of art in multimodal integration based on the SC. A neural computational model of the subcortical Superior Colliculus is being designed to demonstrate the multimodal integration that is performed in the deep layers. The enhancement and depression phenomena with low signal strength are demonstrated and the impact of multimodal integration is discussed. The presented model provides a first insight into the computational modeling and performance of the SC based on the concepts of Stein and Meredith and highlights the generated multimodal output. This work indicates a lot of potential for subsequent research for the model to emerge as a full computational multimodal sensory data integration model of the Superior Colliculus.

Acknowledgements

The authors extend their gratitude to Chris Rowan and Eric Hill for the assistance in setting up the experimental environment.

References

1. Green, A., Eklundh, K.S.: Designing for Learnability in Human-Robot Communication. IEEE Transactions on Industrial Electronics 50(4), 644–650 (2003)
2. Arai, K., Keller, E.L., Edelman, J.A.: A Neural Network Model of Saccade Generation Using Distributed Dynamic Feedback to Superior Colliculus. In: Proceedings of International Joint Conference on Neural Networks, pp. 53–56 (1993)
3. Stein, B.E., Meredith, M.A.: The Merging of the Senses. Cognitive Neuroscience Series. MIT Press, Cambridge (1993)
4. Quaia, C., Lefevre, P., Optican, L.M.: Model of the Control of Saccades by Superior Colliculus and Cerebellum. Journal of Neurophysiology 82(2), 999–1018 (1999)
5. Cuppini, C., Magosso, E., Serino, A., Pellegrino, G.D., Ursino, M.: A Neural Network for the Analysis of Multisensory Integration in the Superior Colliculus. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D.P. (eds.) ICANN 2007, Part II. LNCS, vol. 4669, pp. 9–11. Springer, Heidelberg (2007)

6. Gilbert, C., Kuenen, L.P.S.: Multimodal Integration: Visual Cues Help Odour-Seeking Fruit Flies. Current Biology 18, 295–297 (2008)
7. Fitzpatrick, D.C., Kuwada, S., Batra, R.: Transformations in processing Interaural time difference between the superior olivary complex and inferior colliculus: beyond Jeffress model. Hearing Research 168, 79–89 (2002)
8. Massaro, D.W.: A Framework for Evaluating Multimodal integration by Humans and A Role for Embodied Conversational Agents. In: Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI 2004), pp. 24–31 (2004)
9. Droulez, J., Berthoz, A.: A neural network model of sensoritopic maps with predictive short-term memory properties. In: Proceedings of the National Academy of Sciences, USA, Neurobiology, vol. 88, pp. 9653–9657 (1991)
10. Girard, B., Berthoz, A.: From brainstem to cortex: Computational models of saccade generation circuitry. Progress in Neurobiology 77, 215–251 (2005)
11. Gurney, K.: Integrative computation for autonomous agents: a novel approach based on the vertebrate brain. Talk presented at EPSRC Novel computation initiative meeting (2003)
12. Yavuz, H.: An integrated approach to the conceptual design and development of an intelligent autonomous mobile robot. Robotics and Autonomous Systems 55, 498–512 (2007)
13. Hanheide, M., Bauckhage, C., Sagerer, G.: Combining Environmental Cues & Head Gestures to Interact with Wearable Devices. In: Proceedings of 7th International Conference on Multimodal Interfaces, pp. 25–31 (2005)
14. Laubrock, J., Engbert, R., Kliegl, R.: Fixational eye movements predict the perceived direction of ambiguous apparent motion. Journal of Vision 8(14), 1–17 (2008)
15. Lewald, J., Ehrenstein, W.H., Guski, R.: Spatio-temporal constraints for auditory-visual integration. Behavioural Brain Research 121(1-2), 69–79 (2001)
16. Wolf, J.C., Bugmann, G.: Linking Speech and Gesture in Multimodal Instruction Systems. In: The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2006), UK (2006)
17. Maas, J.F., Spexard, T., Fritsch, J., Wrede, B., Sagerer, G.: BIRON, What's the topic? – A Multi-Modal Topic Tracker for improved Human-Robot Interaction. In: The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2006), UK (2006)
18. Armingol, J.M., del la Escalera, A., Hilario, C., Collado, J.M., Carrasco, J.P., Flores, M.J., Postor, J.M., Rodriguez, F.J.: IVVI: Intelligent vehicle based on visual information. Robotics and Autonomous Systems 55, 904–916 (2007)
19. Pavon, J., Gomez-Sanz, J., Fernandez-Caballero, A., Valencia-Jimenez, J.J.: Development of intelligent multisensor surveillance systems with agents. Robotics and Autonomous Systems 55, 892–903 (2007)
20. Juan, C.H., Muggleton, N.G., Tzeng, O.J.L., Hung, D.L., Cowey, A., Walsh, V.: Segregation of Visual Selection and Saccades in Human. Cerebral Cortex 18(10), 2410–2415 (2008)
21. Groh, J.: Sight, sound processed together and earlier than previously thought, 919-660-1309, Duke University Medical Centre (2007) (Released, 29 October 2007)
22. Jolly, K.G., Ravindran, K.P., Vijayakumar, R., Sreerama Kumar, R.: Intelligent decision making in multi-agent robot soccer system through compounded artificial neural networks. Robotics and Autonomous Systems 55, 589–596 (2006)
23. King, A.J.: The Superior Colliculus. Current Biology 14(9), R335–R338 (2004)
24. Kohonen, T.: Self-Organized formation of Topographical correct feature Maps. Biological Cybernetics 43, 59–69 (1982)

25. Voutsas, K., Adamy, J.: A Biologically inspired spiking neural network for sound source lateralization. IEEE transactions on Neural Networks 18(6), 1785–1799 (2007)
26. Calms, L., Lakemeyer, G., Wagner, H.: Azimuthal sound localization using coincidence of timing across frequency on a robotic platform. Acoustical Society of America 121(4), 2034–2048 (2007)
27. Lee, M., Ban, S.-W., Cho, J.-K., Seo, C.-J., Jung, S.K.: Modeling of Saccadic Movements Using Neural Networks. In: International Joint Conference on Neural Networks, vol. 4, pp. 2386–2389 (1999)
28. Coen, M.H.: Multimodal Integration – A Biological View. In: Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI 2001), pp. 1414–1424 (2001)
29. Beauchamp, M.S., Lee, K.E., Argall, B.D., Martin, A.: Integration of Auditory and Visual Information about Objects in Superior Temporal Sulcus. Neuron 41, 809–823 (2004)
30. Bennewitz, M., Faber, F., Joho, D., Schreiber, M., Behnke, S.: Integrating Vision and Speech for Conversations with Multiple Persons. In: Proceedings of International Conference on Intelligent Robots and System, IROS (2005)
31. Murray, J., Erwin, H., Wermter, S.: A Hybrid Architecture using Cross-Correlation and Recurrent Neural Networks for Acoustic Tracking in Robots. In: Wermter, S., Palm, G., Elshaw, M. (eds.) Biomimetic Neural Learning for Intelligent Robots, pp. 55–73 (2005)
32. Paleari, M., Christine, L.L.: Toward Multimodal Fusion of Affective Cues. In: Proceedings of International Conference on Human Computer Multimodality HCM 2006, pp. 99–108 (2006)
33. Casey, M.C., Pavlou, A.: A Behavioural Model of Sensory Alignment in the Superficial and Deep Layers of the Superior Colliculus. In: Proceeding of International Joint Conference on Neural Networks (IJCNN 2008), pp. 2750–2755 (2008)
34. Mavridis, N., Roy, D.: Grounded Situation Models for Robots: Where words and percepts meets. In: Proceedings of International Conference on Intelligent Robots and Systems (IEEE/RSJ), pp. 4690–4697 (2006)
35. Kubota, N., Nishida, K., Kojima, H.: Perceptual System of A Partner Robot for Natural Communication Restricted by Environments. In: Proceedings of International Conference on Intelligent Robots and Systems (IEEE/RSJ), pp. 1038–1043 (2006)
36. Palanivel, S., Yegnanarayana, B.: Multimodal person authentication using speech, face and visual speech. Computer Vision and Image Understanding (IEEE) 109, 44–55 (2008)
37. Patton, P., Belkacem-Boussaid, K., Anastasio, T.J.: Multimodality in the superior colliculus: an information theoretic analysis. Cognitive Brain Research 14, 10–19 (2002)
38. Pattion, P.E., Anastasio, T.J.: Modeling Cross-Modal Enhancement and Modality-Specific Suppression in Multisensory Neurons. Neural Computation 15, 783–810 (2003)
39. Cucchiara, R.: Multimedia Surveillance Systems. In: 3rd International Workshop on Video Surveillance and Sensor Networks (VSSN 2005), Singapore, pp. 3–10 (2005) ISBN: 1-59593-242-9
40. Rothwell, J.C., Schmidt, R.F.: Experimental Brain Research, vol. 221. Springer, Heidelberg, SSN: 0014-4819
41. Schauer, C., Gross, H.M.: Design and Optimization of Amari Neural Fields for Early Auditory – Visual Integration. In: Proc. Int. Joint Conference on Neural Networks (IJCNN), Budapest, pp. 2523–2528 (2004)
42. Stiefelhagen, R.: Tracking focus of attention in meetings. In: International conference on Multimodal Interfaces (IEEE), Pittsburgh, PA, pp. 273–280 (2002)
43. Stiefelhagen, R., Bernardin, K., Ekenel, H.K., McDonough, J., Nickel, K., Voit, M., Wolfel, M.: Auditory-visual perception of a lecturer in a smart seminar room. Signal Processing 86, 3518–3533 (2006)

44. Wermter, S., Weber, C., Elshaw, M., Panchev, C., Erwin, H., Pulvermuller, F.: Towards multimodal neural robot learning. Robotics and Autonomous Systems 47, 171–175 (2004)
45. Stork, D.G., Wolff, G., Levine, E.: Neural Network lip reading system for improved speech recognition. In: Proc. Intl. Conf. Neural Networks (IJCNN 1992), vol. 2, pp. 289–295 (1992)
46. Steil, J.J., Rothling, F., Haschke, R., Ritter, H.: Situated robot learning for multi-modal instruction and imitation of grasping. Robotics and Autonomous Systems 47, 129–141 (2004)
47. Huwel, S., Wrede, B., Sagerer, G.: Robust Speech Understanding for Multi-Modal Human-Robot Communication. In: 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2006), UK, pp. 45–50 (2006)
48. Trappenberg, T.: A Model of the Superior Colliculus with Competing and Spiking Neurons., BSIS Technical Report, No. 98-3 (1998)
49. Anastasio, T.J., Patton, P.E., Belkacem-Baussaid, K.: Using Bayes' Rule to Model Multisensory Enhancement in the Superior Colliculus. Neural Computation 12, 1165–1187 (2000)
50. Perrault Jr., T.J., William Vaughan, J., Stein, B.E., Wallace, M.T.: Superior Colliculus Neurons use Distinct Operational Modes in the Integration of Multisensory Stimuli. Journal of Neurophysiology 93, 2575–2586 (2005)
51. Stanford, T.R., Stein, B.E., Quessy, S.: Evaluating the Operations Underlying Multisensory Integration in the Cat Superior Colliculus. The Journal of Neuroscience 25(28), 6499–6508 (2005)
52. Spexard, T., Li, S., Booij, O., Zivkovic, Z.: BIRON, where are you?—Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization. In: Proceedings of International Conference on Intelligent Robots and Systems (IEEE/RSJ), pp. 934–940 (2006)
53. Trifa, V.M., Koene, A., Moren, J., Cheng, G.: Real-time acoustic source localization in noisy environments for human-robot multimodal interaction. In: Proceedings of RO-MAN 2007 (IEEE International Symposium on Robot & Human Interactive Communication), Korea, pp. 393–398 (2007)
54. Cutsuridis, V., Smyrnis, N., Evdokimidis, I., Perantonis, S.: A Neural Model of Decision-making by the Superior Colliculus in an Anti-saccade task. Neural Networks 20, 690–704 (2007)
55. Wallace, M.T., Meredith, M.A., Stein, B.E.: Multisensory Integration in the Superior Colliculus of the Alert Cat. Journal of Neurophysiology 80, 1006–1010 (1998)
56. Wilhelm, T., Bohme, H.J., Gross, H.M.: A Multi-modal system for tracking and analyzing faces on a mobile robot. Robotics and Autonomous Systems 48, 31–40 (2004)
57. Zou, X., Bhanu, B.: Tracking humans using Multi-modal Fusion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005), p. 4 (2005)