A Computational Model of Crossmodal Processing for Conflict Resolution

German I. Parisi[1], Pablo Barros[1], Matthias Kerzel[1], Haiyan Wu[2,3], Guochun Yang[2,3], Zhenghan Li[2,3], Xun Liu[2,3], Stefan Wermter[1]
[1]Knowledge Technology, Department of Informatics, University of Hamburg, Germany
[2]CAS Key Laboratory of Behavioral Science, Chinese Academy of Sciences, Beijing, China
[3]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

***Abstract*—The brain integrates information from multiple sensory modalities to form a coherent and robust perceptual experience in complex environments. This ability is progressively acquired and fine-tuned during developmental stages in a multisensory environment. A rich set of neural mechanisms supports the integration and segregation of multimodal stimuli, providing the means to efficiently solve conflicts across modalities. Therefore, there is the motivation to develop efficient mechanisms for robotic platforms that process multisensory signals and trigger robust sensory-driven motor behavior. In this paper, we implement a computational model of crossmodal integration in a sound source localization task that accounts also for audiovisual conflict resolution. Our model consists of two layers of reciprocally connected visual and auditory neurons and a layer with crossmodal neurons that learns to integrate (or segregate) audiovisual stimuli on the basis of spatial disparity. To validate our architecture, we propose a spatial localization task in which 30 subjects had to determine the location of the sound source in a virtual scenario with four animated avatars. We measured their accuracy and reaction time under different conditions for congruent and incongruent audiovisual stimuli. We used this study as a baseline to model human-like behavioral responses with a neural network architecture exposed to the same experimental conditions.**

I. Introduction

The ability of the brain to integrate information conveyed by multiple sensory sources is crucial for the efficient interaction with the environment [1]. A vast body of behavioral studies has reported on a large set of phenomena showing the ability of humans to perceptually integrate multisensory information, e.g., localizing objects and events from audiovisual stimuli such as light blobs and sound clicks [2, 3].

A widely studied effect of multimodal integration is the ventriloquist illusion, which refers to perceiving sounds as coming from a different location than their actual location due to perception being strongly biased by visual stimuli [4, 5]. This form of audiovisual integration has been argued to be the result of a near-optimal bi-modal integration strategy in the brain for stimuli exhibiting a small spatial or temporal disparity [6]. Multimodal processing also involves the segregation of sensory inputs that are assumed not to be caused by the same source [7]. Audiovisual stimuli exhibiting a lower stimulus-response compatibility are expected to have less influence on each other. In psychology, audiovisual conflicts have been commonly studied in the context of crossmodal processing [8]. For instance, it has been shown that increasing spatial disparity leads to decreasing visual bias on auditory localization [2, 3, 5]. However, the problem of inferring whether a bi-modal stimulus is caused by a common source has been shown to depend on a larger number of factors such as spatial and temporal congruency, prior knowledge, and expectations [9].

From a neurophysiological perspective, a large number of brain areas have been associated with the processing of multisensory information. An example is the superior colliculus (SC), a subcortical area of the mammalian brain that exhibits multimodal behavior for target selection and producing reflexive motor responses such as head-eye movements [1]. Neurons selective to complex audiovisual patterns have been found, e.g., in the superior temporal sulcus (STS) which is argued to link unimodal representations from cortical areas (visual and auditory cortex) and to account for the association of highly correlated visual and linguistic stimuli [10]. Cortical and subcortical areas are known to interact resulting in perception and behavior being driven by the interplay of low-level sensory stimuli and higher-order spatial-semantic cues [3, 9]. While low-level stimuli may be integrated or segregated on the basis of their spatial and temporal alignment, the experience-driven development of internal representations in associative areas of the brain modulates multimodal interaction on the basis of semantic congruence. Thus, our ability to integrate multimodal stimuli and solve crossmodal conflicts is progressively acquired and fine-tuned through the exposure to multimodal events, with critical periods during early developmental stages playing a crucial role for cortical and subcortical organization [11]. This learning process endows the brain with the ability to better adapt to difficult perceptual conditions met during our daily experience, e.g., weak or noisy external stimuli.

Artificial systems embedded with multimodal processing capabilities may result in a more robust perceptual experience [12], especially in the case of sensory uncertainty [13]. The task of causal inference is crucial for triggering sensory-driven motor behavior, e.g., computing a single spatial position from audiovisual input to produce eye-head movements. However, learning models that address multisensory causal inference for the resolution of conflicts have remained an open issue for artificial systems and robots. Neural network models have been proposed that implement Bayesian inference

978-1-5386-3715-9/17/$31.00 ©2017 IEEE

principles fitting behavioral data from different multimodal tasks (see [14] for a recent review). In particular, it has been shown that two layers of unisensory neurons with reciprocal connections trained on multimodal data are sufficient to account for the ventriloquist effect [15]. Nevertheless, the problem of computing one single cause (position) to drive behavior requires an additional layer of crossmodal neurons that process the output of the two unisensory layers [16].

In this paper, we implement a computational model that accounts for audiovisual conflict resolution. In line with recent neural network architectures for multimodal integration and conflict resolution [15, 16], our learning model consists of three layers: two upstream layers of visual and auditory neurons, and a downstream layer with crossmodal neurons that learns to integrate (or segregate) audiovisual stimuli. For the validation of our model, as an extension to previous studies conducted on light blobs and sound clicks [2, 3], we propose a spatial localization task in a virtual scenario with four animated avatars with lip movement as visual patterns and spoken words as auditory patterns. We conducted a behavioral study with 30 subjects to evaluate their accuracy and reaction time in localizing the sound source under different conditions for congruent and incongruent audiovisual patterns. Consistent with previous studies showing the effects of visual bias over auditory stimuli for simple audiovisual stimuli, the analysis of the data from our four-figure scenario resulted in a decreased localization accuracy and higher reaction time for incongruent audiovisual patterns. Experimental results show that our model accounts for fitting behavioral data from the proposed spatial localization task, thus providing a prominent baseline for triggering human-like responses on a humanoid robot exposed to similar conditions.

II. Behavioral Study

Thirty volunteers participated in the experiment (14 males and 16 females, aged between 17 and 30, and right-handed). All participants declared normal or corrected-to-normal hearing and visual acuity and no history of neurological or psychiatric disorder. The task consisted of a spatial location task in which the participants have to choose the source of the sound given a set of congruent and incongruent audiovisual patterns. Two of them (one male and one female) were excluded from further analysis for the low accuracy (29% and 30% respectively), which was lower than the 95% confidence interval of the sample (30.6%-82.4%). We suppose the low accuracy may be caused by the inability of normally discriminating audiovisual location. The other 28 participants were in the range 38%-79% accuracy (mean: 56%, standard deviation: 13%).

A. Apparatus, Stimuli and Procedure

We used the four-figure scenario background (Fig. 1) with videos containing lip movement for one of the figures as visual stimuli and short words (i.e. ”la”, ”wa”, ”ha”) from a synthesized voice as auditory stimuli. The visual stimuli were displayed on a 17-inch LCD monitor with the viewing distance of approximately 60 cm and the spoken words were presented via a headphone set.

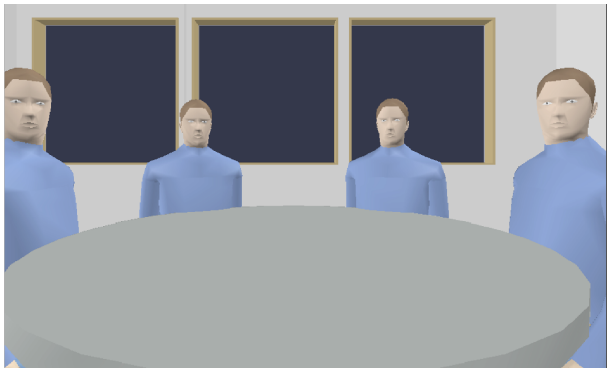

Fig. 1. Illustration of the scenario with four virtual avatars sitting around a table. Lip movements and short spoken words from a synthesized voice are used as audiovisual stimuli.

The visual stimuli were generated in a virtual environment (VREP)[1] using a set of modified human avatars. Different 3D meshes for the face area (with varying jaw angles) were used during recording to create the impression of lip and mouth movement during speech. The videos were rendered with OpenGL with a 1024 $\times$ 1024 pixel resolution, and the virtual camera had an 80° opening angle (Fig. 2). Videos were recorded from the screen and cut to 1-second clips for each of the four figures moving and not moving its lips. The binaural auditory stimuli were synthesized with two head-related transfer functions (HRTFs) for the left and right audio channel. These functions emulate the temporal and the level differences of perceived sounds by the left and right ear due to different traveling times and damping when a sound comes from different points in space.

The task started with a static four-figure background for 500 ms, followed by a video with lip movement for 1000 ms, and then another 500 ms of static background. Only one figure would move its lips during each trial, which conveys the visual spatial information with four possible locations: Left2 (L2), Left1 (L1), Right1 (R1) and Right2 (R2). The visual angle between each two-figure pairs from the participants' perspective was about 10°. The pitches of the two auditory channels were edited so that the sounds were perceived with spatial location from different angles: -60° (L2), -20° (L1), 20° (R1), and 60° (R2), with these angles being amplified to create more distinct auditory stimuli.

We created a set of 16 (4×4) conditions which were further classified into 5 combined conditions as follows:

- **Condition 0**: congruent condition, such as visual L1-audio L1 (L1L1);
- **Condition 1a**: conflict occurred on the same side, such as L2L1;
- **Condition 1b**: conflict occurred between the central two locations, such as L1R1;

[1]VREP - http://www.coppeliarobotics.com/

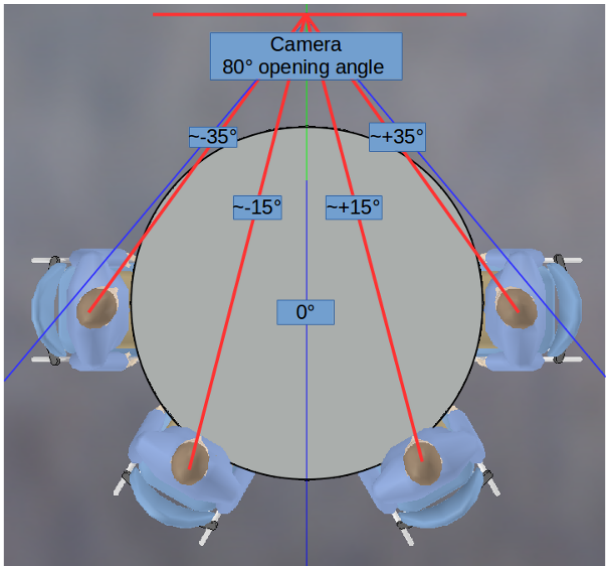

Fig. 2. Illustration of the rendered video in our virtual environment (VREP). It is possible to see the camera angles for each virtual avatar.

- **Condition 2**: conflict occurred between the locations with a one-figure interval, such as L2R1;
- **Condition 3**: conflict occurred between the locations with a two-figure interval, such as L2R2.

We expect to observe increased congruency effects in Condition 1 (including 1a and 1b) with respect to Condition 3.

Participants first took part in two pre-task studies with 10 trials each to become familiar with the auditory stimuli and the cross-modal scenario. The formal test consisted of four blocks, each containing 120 trials of randomized visual-auditory combinations. We asked the subjects to produce a four-key response to the locations of sound regardless of the lip movement. During the task, they were instructed to watch the screen attentively to ensure the effectiveness of visual information.

B. Data Analysis

The obtained behavioral data were analyzed with dependent variables of both the reaction time (RT) and the error rate (ER) of localization. Error trials and trials with RT beyond three standard deviations were excluded. We conducted one-way repeated measures analysis of variance (ANOVA). Results are shown in Fig. 3. A significant difference among the five conditions was observed in both RT ($F(4, 108) = 18.478$, $p < 0.001$, $\eta_p^2 = 0.406$) and ER ($F(4, 108) = 9.477$, $p < 0.001$, $\eta_p^2 = 0.260$).

In comparison with Condition 0, subjects responded with significantly higher ER in Condition 1a (45%, $p < 0.001$), Condition 1b (46%, $p < 0.01$) and Condition 2 (45%, $p < 0.001$). In comparison with Condition 3 (ER=33%), participants responded with significantly higher ER in Condition 1a ($p < 0.05$) and Condition 2 ($p < 0.01$) respectively.

In comparison with Condition 0 (RT=935 ms), participants responded slower in Condition 1a (980 ms, $p < 0.01$), Condition 1b (1066 ms, $p < 0.001$) and Condition 2 (1009 ms, $p < 0.001$) respectively. In comparison with Condition 3, participants responded slower in Condition 1a ($p < 0.05$), Condition 1b ($p < 0.001$) and Condition 2 ($p < 0.05$) respectively. Moreover, Condition 1b shows a slower response than Condition 1a ($p < 0.01$).

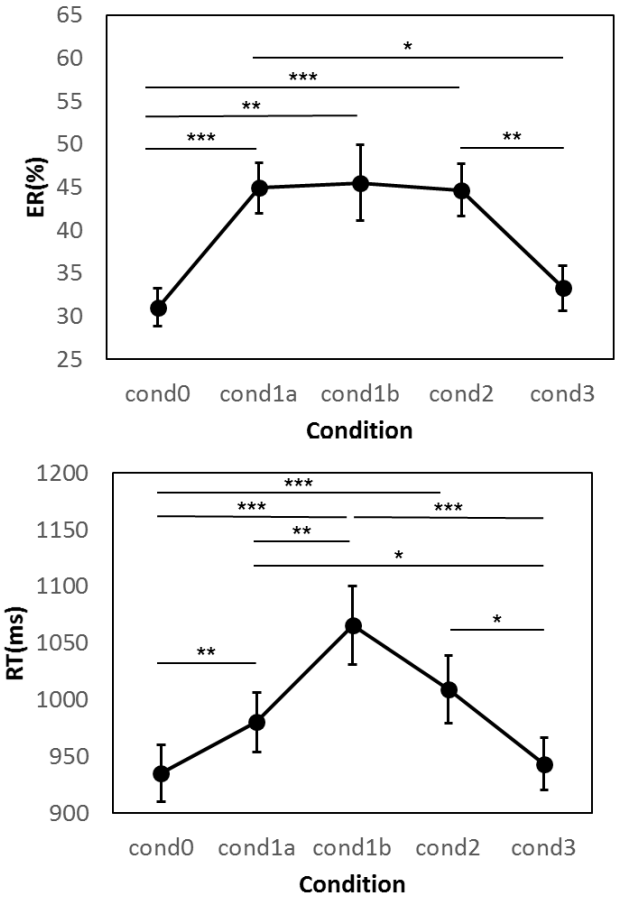

Fig. 3. Stimulus-Response Conflict effect during the sound source localization task from audiovisual stimuli in terms of error rate (ER, upper plot) and reaction time (RT, bottom plot).

C. Discussion

In terms of accuracy, the data suggest that the task of selecting the correct target audio location is disrupted by the visual distractor. This result is consistent with behavioral studies using simpler audiovisual stimuli such as light blobs and clicks (e.g. [3]), in which a small spatial disparity between the visual and auditory stimuli leads to a strong visual bias that shifts the perception of the auditory stimulus towards the visual one (i.e. the ventriloquism effect [4]). Therefore, although our scenario conveys more semantics with respect to light blobs and clicks, subjects were much more inaccurate to locate the position of the sound from the exposure to incongruent lip-word pairs. In this context, additional studies are required to measure whether a scene conveying more semantics (e.g., more realistic and animated avatars with human voices) leads to attenuate the ventriloquism effect, e.g., due to the bias

of semantics and prior expectations modulating the low-level localization task.

In terms of reaction time, the responses obtained in this study are generally slower (about 1000 ms) compared to our previous conflict experiments (about 400-600 ms) [17, 18], suggesting that this task is relatively difficult. One possible reason is that the subjects needed to produce a response by using four keys for different locations in the current study, which leads to a longer reaction time. We found robust conflict effects in Condition 1a, 1b and 2 but not in Condition 3 (although Condition 3 contains the largest spatial inconsistency). It is possible that the conflict could be well resolved when the interference information comes from a location relative far from the actual one, as there is enough time to produce a response. Thus, the attention was shifted back to the sound location from the inconsistent visual cue before a decision was made. Furthermore, although conflicts in Condition 1a and 1b both occurred between adjacent locations with the same auditory-visual angle (40°), we observed a significantly slower reaction time in Condition 1b, indicating that it is more difficult for people to discriminate the sound location at the center than at the lateral sides. This is consistent with the fact that visual spatial resolution is higher at the center while the auditory resolution is higher at the sides [19].

III. Computational Model

The computational model learns to integrate (or segregate) audiovisual input from the exposure to a set of unisensory and multisensory stimuli during a training session. The neural architecture consists of two upstream layers of N visual and N auditory neurons and a downstream layer with N crossmodal neurons (Fig. 4). This architecture is based on the two-layer architecture proposed by Magosso *et al.* [15] for audiovisual integration, extended with a third layer for the causal inference problem [16]. Neurons are topologically aligned and each neuron codes for a specific position of space. We set $N = 180$ so that the distance between each neuron is 1°, covering an area of 180° in the visual and the auditory space.

A. Audiovisual Input and Neural Receptive Fields

The visual and auditory input is represented as Gaussian functions resembling spatially localized external stimuli filtered by the receptive fields of unisensory neurons. The mean of the Gaussian (p^v and p^a for the visual and the auditory modality respectively) corresponds to the position of the stimulus in the external world, while the standard deviation (σ^v, σ^a) corresponds to the width of the receptive fields of the neurons. The model assumes that the auditory and visual area are spatially organized, with the spatial resolution of auditory neurons being smaller than the spatial resolution of visual input. This difference in the spatial resolution of the auditory and visual neurons is introduced by setting $\sigma^a < \sigma^v$. The output activity of neurons is computed from the weighted sum of its inputs and normalized to 0 and 1, i.e., 1 is the maximum activity.

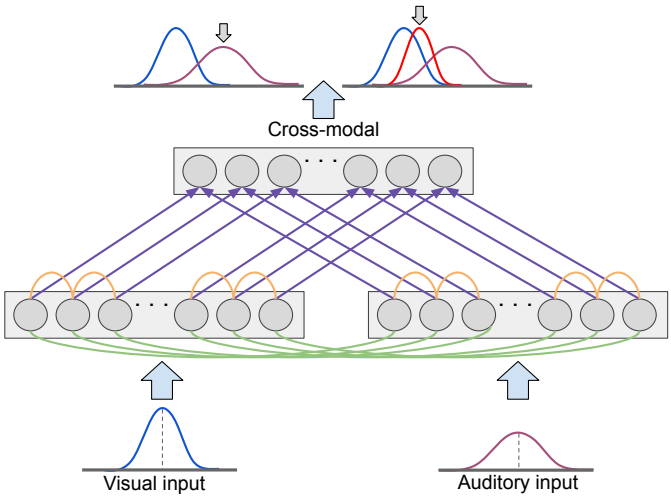

Fig. 4. Neural network architecture for multimodal integration. The architecture consists of two upstream layers of visual and auditory neurons and a downstream layer with crossmodal neurons for the causal inference problem. Neurons are topologically aligned and each neuron codes for a specific position of space. Each unisensory layer comprises lateral connections (orange lines) and reciprocal inter-layer connections (green lines). Neurons in the crossmodal layer receive input from the two unimodal neurons located at the same spatial position (purple arrows).

Neurons in each unimodal layer are connected through lateral synapses that include both excitatory and inhibitory effects (Fig. 4, orange lines). The neurons are arranged with a Mexican-hat disposition to excite proximal neurons and inhibit distal neurons, yielding a competitive mechanism between stimuli within a unisensory layer.

B. Multimodal Processing

Neurons in the two unimodal layers are reciprocally connected through inter-area excitatory synapses (Fig. 4, green lines), so that each neuron receives input only from the neuron of the other modality at the same spatial position. These connections modulate the influence of one modality over the other. Therefore, each output $u_j^m(t)$ of a neuron within a modality m (with $m = v$ or $m = a$) processes input as the sum of the external input $e_j^m(t)$, the intra-area lateral input $l_j^m(t)$, and the crossmodal the from inter-layer input $c_j^m(t)$.

Different to the intra-layer connections defining neural receptive fields and that we assume to be pre-defined, we train crossmodal connections with a percentage of unimodal and multimodal input. The learning of connectivity patterns is carried out via Hebbian training rules for synaptic potentiation. It has been shown that the Hebbian-like development of inter-layer connections accounts for the ventriloquism effect, where the perception of the auditory stimulus is shifted in the direction of the visual one provided that the spatial discrepancy between the two stimuli is smaller than 20-25 degrees [15]. Conversely, when the spatial discrepancy is higher, the effect of the integration of the two stimuli is negligible. This behavior is consistent with a Bayesian estimator that sub-optimally computes the prior and likelihood probabilities for inferring the position of multimodal stimuli [16].

Each neuron provides a two-dimensional vector according to the population vector metric, i.e. its length is equal to the firing rate and the phase equal to twice its label. The output z^m_{vet} of the model ($m = v, a$) is the perceived stimulus location, summing up all vectors such that

$$z^m_{vet} = \frac{1}{2}\arctan\left(\frac{\sum_{k=1}^{N} y^m_k \cdot \sin(2k)}{\sum_{k=1}^{N} y^m_k \cdot \cos(2k)}\right), \quad (1)$$

where y^m_k represents the activity of the neuron at position k.

In the crossmodal layer, a single position is computed using the maximum activity within this layer. Each crossmodal neuron receives input only from the two unimodal neurons located at the same spatial position via identical weights (Fig. 4, purple arrows), in line with the assumption that the sum of population codes account for the optimal Bayesian inference [16]. The output of the crossmodal neurons is computed as:

$$y^c_k = \phi\left(w^{ca} y^a_k + w^{cv} y^v_k\right), \quad (2)$$

where y^c_k is the activity of the crossmodal neuron at position k, w^{cm} are the synapse weights from the neurons of modality m to the crossmodal layer, and ϕ is a monotonically decreasing function.

C. Experimental Results

We trained the neural architecture to a basal state and used it to test stimulus-response conflicts for our spatial localization task. The architecture was trained with a set of unimodal and congruent-incongruent multimodal stimuli containing all possible audiovisual spatial combinations (for 180 possible positions for each layer). We set equal input strength for the visual and auditory input ($E^m_0 = 15$) with $\sigma^v = 4°$ and $\sigma^a = 32°$, thus yielding a higher spatial resolution for the visual modality. Training sessions were conducted assuming a set of model parameters reported in [15, 16]. For testing the basal networks, we created input reproducing the experimental conditions from our behavioral study, i.e. comprising 16 (4×4) congruent-incongruent audiovisual pairs (see Section II.A). We validated the architecture by computing the accuracy for the localization of the auditory target. As can be seen in Fig. 5, the architecture reproduces behavioral measures for the localization task in terms of accuracy (Fig. 3).

Responses from the model were instantaneous, yielding a very similar reaction time (RT) for congruent and incongruent stimuli. In this case, a comparison between the model's RT and the one from the subjects is not justified since higher RT in humans may be caused by distinct neural processing pathways (see Section II.C for discussion), whereas our architecture does not consider this aspect.

For simplicity, we trained inter-layer connections and crossmodal connections from the two unisensory layers, whereas lateral connections defining the unisensory receptive fields were pre-defined. Furthermore, for a more biologically realistic experiment, the exposure of the networks to persistent multimodal stimuli should modify connectivity patterns, thereby accounting for the ventriloquism aftereffect, i.e. the

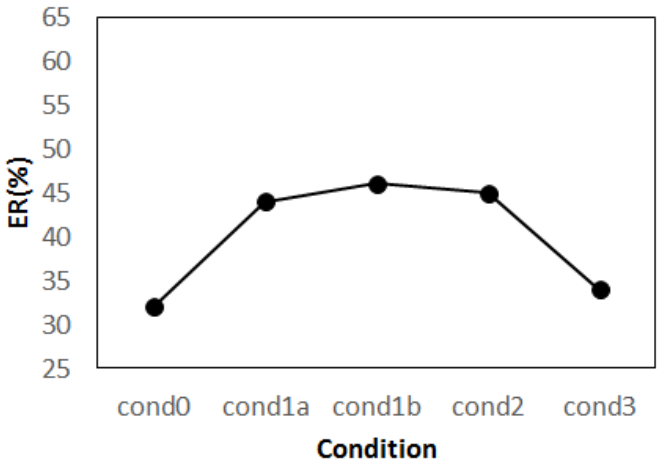

Fig. 5. Stimulus-Response Conflict effect during the sound source localization task exposing the trained neural architecture to audiovisual stimuli.

responses to auditory stimuli are shifted towards the previously presented visual stimulus [20]. In terms of the neural network architecture, this can be seen as the on-line adaptation of the connections between unisensory layers, which is possible by training further inter-layer synapses via Hebbian learning [15].

To be noted is that we have focused on the spatial effects of multimodal integration. Nevertheless, it is known that the temporal component plays a very important role. It has been shown that the current neural network implementation can be extended to account for the spatiotemporal processing of multimodal stimuli [21].

IV. Conclusion and Future Work

In this paper, we presented a neural network model for crossmodal processing and conflict resolution. The neural architecture learns to integrate or segregate audiovisual input from the exposure to a set of unisensory and multisensory training stimuli. For validating our computational model, we conducted a behavioral study on a sound source localization task exposing the subjects to congruent and incongruent audiovisual patterns. As an extension to previous studies using light blobs and sound clicks, we proposed a scenario composed of four animated avatars using lip movements as visual patterns and spoken short words as the auditory ones. In line with these studies using simple stimuli, our data suggest that subjects were much more inaccurate to locate the position of the sound from the exposure to incongruent lip-word pairs. Conditions of conflict have also exhibited a higher reaction time. We hypothesize that this may be caused by incongruent conditions leading to a longer processing pipeline (in terms of neural pathways) for integrating or segregating stimuli on the basis of available scene semantics and knowledge-driven expectations.

The obtained results motivate further research in the direction of behavioral data collection and the neural network architecture. From a behavioral perspective, it would be interesting to test whether a more complex scene, e.g., animated avatars conveying additional information in terms of gender,

face expressions, and body motion, have a significant impact on the localization accuracy and reaction time under congruent and incongruent conditions. Although it is known that prior knowledge and expectations modulate multimodal processing [9], behavioral studies have so far mostly focused on the processing of simple stimuli that neglect the role of cortical areas (e.g., visual and auditory cortex) and higher-level brain areas (e.g., the STS) for the modulation of low-level crossmodal processing (as in the SC). From a neural network perspective, this extension would require the interaction between subcortical and cortical layers that model the interplay of bottom-up and top-down crossmodal modulation. The learning and the recognition of meaningful visual and auditory features from complex audiovisual patterns can be implemented in terms of a hierarchy of neural networks that tune internal representations to process features with an increasing degree of complexity and abstraction.

While the underlying mechanisms of the brain for crossmodal conflict resolution are still to be fully investigated, our work may be seen as a basis for the development of complex artificial systems aimed at triggering human-like behavioral responses driven by multimodal perception. In this context, we argue that the interplay of behavioral studies, neurophysiological findings, and neural network models is crucial for achieving such a goal.

ACKNOWLEDGMENT

This research was supported by National Natural Science Foundation of China (NSFC) and the German Research Foundation (DFG) under project Transregio Crossmodal Learning (TRR 169). The authors would like to thank Josip Josifovski for valuable technical help during the development of the four-figure scenario and Stefan Heinrich and Sascha Griffiths for a revision of early versions of the paper.

REFERENCES

[1] B. E. Stein and M. A. Meredith, *The merging of the senses*. Cambridge, MA, US: The MIT Press, 1993.

[2] K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams, "Causal inference in multisensory perception," *PLoS ONE*, vol. 2, no. 9, 2007.

[3] B. Odegaard, D. Wozny, and L. Shams, "Biases in visual, auditory, and audiovisual perception of space.," *PLoS Computational Biology*, vol. 11, no. 12, 2015.

[4] C. E. Jack and W. R. Thurlow, "Effects of degree of visual association and angle of displacement on the "ventriloquism" effect," *Perceptual & Motor Skills*, vol. 37, pp. 967–979, 1973.

[5] M. Radeau and P. Bertelson, "Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations," *Perception & Psychophysics*, vol. 22, no. 2, pp. 137–146, 1977.

[6] I. B. Witten and E. I. Knudsen, "Why seeing is believing: Merging auditory and visual worlds," *Neuron*, vol. 48, no. 3, 2005.

[7] W.-H. Zhang, H. Wang, K. Y. M. Wong, and S. Wu, "Congruent and opposite neurons: Sisters for multisensory integration and segregation," in *NIPS 29*, pp. 3180–3188, 2016.

[8] P. Zhou and X. Liu, "Attentional modulation of emotional conflict processing with flanker tasks," *PLoS One*, vol. 8, no. 3, 2013.

[9] C. Kayser and L. Shams, "Multisensory causal inference in the brain," *PLoS Biology*, vol. 13, no. 2, 2015.

[10] N. E. Barraclough, D. Xiao, C. I. Baker, M. W. Oram, and D. I. Perrett, "Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions.," *Journal of Cognitive Neuroscience*, vol. 17, no. 3, pp. 377–391, 2005.

[11] T. K. Hensch, "Critical period plasticity in local cortical circuits.," *Nature Reviews*, vol. 6, pp. 877–888, 2005.

[12] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks.," *Robotics and Autonomous Systems*, vol. 62, no. 6, pp. 721 – 736, 2014.

[13] F. Cruz, G. Parisi, J. Twiefel, and S. Wermter, "Multi-modal integration of dynamic audiovisual patterns for an interactive reinforcement learning scenario," in *Proceedings of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 759–766, 2016.

[14] M. Ursino, C. Cuppini, and E. Magosso, "Neurocomputational approaches to modelling multisensory integration in the brain: A review," *Neural Networks*, vol. 60, pp. 141 – 165, 2014.

[15] E. Magosso, C. Cuppini, and M. Ursino, "A neural network model of ventriloquism effect and aftereffect," *PLoS ONE*, vol. 7, no. 8, 2012.

[16] M. Ursino, C. Cuppini, and E. Magosso, "Multisensory bayesian inference depends on synapse maturation during training: Theoretical analysis and neural modeling implementation," *Neural Computation*, vol. 29, no. 3, pp. 735–782, 2017.

[17] Q. Li, W. Nan, K. Wang, and X. Liu, "Independent processing of stimulus-stimulus and stimulus-response conflicts," *PLoS ONE*, vol. 9, 02 2014.

[18] X. Liu, Y. Park, X. Gu, and J. Fan, "Dimensional overlap accounts for independence and integration of stimulus—response compatibility effects," *Attention, Perception, & Psychophysics*, vol. 72, no. 6, pp. 1710–1720, 2010.

[19] W. D. Hairston, M. T. Wallace, J. W. Vaughan, B. E. Stein, J. L. Norris, and J. A. Schirillo, "Visual localization ability influences cross-modal bias," *Journal of Cognitive Neuroscience*, vol. 15, no. 1, pp. 20–29, 2003.

[20] P. Bertelson, I. Frissen, J. Vroomen, and B. de Gelder, "The aftereffects of ventriloquism: patterns of spatial generalization.," *Perception Psychology*, vol. 68, pp. 428–436, 2006.

[21] C. Cuppini, E. Magosso, N. Bolognini, G. Vallar, and M. Ursino, "A neurocomputational analysis of the sound-induced flash illusion.," *Neuroimage*, vol. 15, no. 92, pp. 248–266, 2014.