A Neurorobotic Experiment for Crossmodal Conflict Resolution

German I. Parisi[1], Pablo Barros[1], Di Fu[1,2,3], Sven Magg[1], Haiyan Wu[2,3,4], 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 (CAS), Beijing, China
[3]Department of Psychology, University of CAS, Beijing, China
[4]Division of the Humanities and Social Sciences, Caltech, CA, USA
http://github.com/knowledgetechnologyuhh/CML_A5_Neurorobotic_IROS2018

1. Introduction

Robots operating in the real world must efficiently interact with their surroundings. Similarly to biological agents, robots can make use of an array of sensors for processing multiple modalities such as vision and audio with the goal of promptly undertaking behaviourally relevant decisions through the combination of sensory observations with prior knowledge and expectations (e.g. internally generated models of the world) [6]. The integration of multisensory information has been widely studied in the literature, e.g., for the learning of multisensory representations from heterogeneous sensor data [12]. However, it is typically assumed that the multisensory measurements provide a complete and coherent data stream that can be integrated on the basis of spatial and/or temporal coincidence. Nevertheless, behaviour should be swift and singular also in situations of sensory uncertainty and multisensory conflict [11].

A vast number of behavioural studies have provided valuable insights into how spatially congruent and incongruent multisensory stimuli can be modelled in artificial systems (e.g., [2]). However, the stimuli used for triggering responses do not reflect the complexity of the environment that artificial agents are expected to interact with. Critically, audio-visual spatial tasks typically use (over)simplified stimuli such as light blobs and sound clicks, and show only one stimulus per modality. Under these conditions, subjects produce responses based on the spatiotemporal congruency of the audio-visual cues but neglect likewise important factors such as semantic congruency and expectations. Such top-down factors significantly contribute to the development of a robust percept [14] and are crucial for modelling multisensory integration in robots.

In this extended abstract, we present a recently published study of crossmodal conflict resolution in a complex environment [9]. To better understand how humans solve crossmodal conflicts, we extended a previously proposed behavioural study with an audio-visual spatial localization task [10]. Our novel study was conducted in an immersive projection environment and comprises a scene with four animated avatars sitting around a table which can produce congruent and incongruent audio-visual stimuli. We trained a deep learning model to trigger human-like responses and evaluated this approach with an iCub robot exposed to the same experimental conditions as human subjects.

Our neurorobotic study contributes to the leverage of current models of robot perception and behaviour taking into account the complex nature of crossmodal environments and the way humans perceive, learn, and act on the basis of rich (and often uncertain) streams of multisensory input. The main contribution of this work is twofold. First, we provide a quantitative analysis of visually-induced bias on the estimation of sound source localization for different types of audio-visual conflicts. Our findings suggest that i) semantics embedded in the scene modulate the magnitude and extension of the visually-induced bias and ii) expectation-driven perceptual mechanisms introduced by the exposure to animated avatars induces a systematic error in the responses comprising static avatars. Second, we implement a deep neural network architecture that models human-like behaviour and triggers similar responses with an iCub robot in real time. The model is motivated by neuroscientific findings suggesting i) the processing of auditory cues (sound source localization) and visual cues (face and body motion) in distinct brain areas and ii) their combination, in terms of neurons responding to (in)congruent multisensory representations, in higher-level areas [1].

2. Behavioural Study

In a previous study [10], we proposed an audio-visual (AV) spatial localization task that comprised a set of 4 animated avatars. The AV stimuli consisted of one avatar with moving lips along with a synchronous, spatially congruent or incongruent auditory cue. Our findings suggest that human subjects were more inaccurate to spatially localize the

Figure 1: Behavioural study on audio-visual localization: (a) Immersive experimental setup with acoustically transparent concave projection screen (b) Subject selecting a position via a keyboard (c) Schematic illustration of one trial of the AV localization task.

sound when exposed to incongruent AV stimuli. However, the study was subject to a number of limitations. First, we tested subjects on AV stimuli comprising one visual and one auditory cue. Crucially, natural scenes may include multiple visual cues influencing multisensory integration to different extents. Therefore, it is important to assess the interplay of multiple visual cues conveying different semantic meaning, e.g., lip and body movement. Second, the visual stimuli were displayed on a 17-inch monitor and the auditory ones were presented via a headphone set, thus significantly differing from natural crossmodal environments and the way humans (and robots) interact with their surroundings. In this novel study, we extend our experimental design to test new hypotheses and propose a new immersive experimental setup so that human subjects and an iCub humanoid robot can be exposed to the same experimental conditions.

For a schematic illustration of our behavioural study, see Fig. 1. A total of 33 subjects (7 female, aged 21–32, right-handed) participated in our experiment. All participants reported that they did not have a history of any neurological conditions (seizures, epilepsy, stroke), and had normal or corrected-to-normal vision and hearing. The AV localization task consisted of the subjects having to select which avatar (out of the 4 avatars in the scene) they believe the auditory cue is coming from. The 4 avatars may move their lips and/or arm in temporal correspondence with an auditory cue. The latter consists of a vocalized combination of 3 syllables (all permutations without repetition composed of ”ha”, ”wa”, ”ba”). The duration of both visual and auditory stimuli is 1000 ms. The experiment comprised 5 AV conditions:

1. **Baseline**: Auditory cue and static avatars.

2. **Moving Lips**: Auditory cue and one avatar with moving lips.

3. **Moving Arm**: Auditory cue and one avatar with a moving arm.

4. **Moving Lips+Arm**: Auditory cue and one avatar with moving lips and arm.

5. **Moving Lips–Arm**: Auditory cue and one avatar with moving lips and another avatar with a moving arm.

For all the conditions except for **Condition 1**, the AV pair may be spatially congruent or incongruent. In **Condition 5**, spatial congruency comprises lips-audio or arm-audio pairs. If we consider all the AV-pair combinations derived from the 5 conditions, it results in 200 trials.

We analyzed our behavioural data in terms of the error rate (ER) with respect to the ground-truth position of the auditory cue. The amount of visually-induced bias on auditory cues depends on the proximity of the cues and their position with respect to the field of view. In the spatial ventriloquism effect, the perception of the auditory stimulus is shifted towards the direction of the visual cue in relation to their spatial proximity. This integration window, however, breaks down when the distance between the two stimuli is greater than 20-25 degrees and the magnitude of the visual bias becomes negligible [4]. Furthermore, visual spatial resolution is higher in the center of the field of view (FOV), thus the magnitude of the visual bias is expected to be higher towards the center rather than towards the periphery [7].

Our findings suggest that the embedded semantics significantly modulate the magnitude and extension (in terms of integration windows) of the visually-induced bias. Moving lips cause higher error rates in the final estimate of the location of the sound with respect to a moving arm (which is visually more salient). This is in line with fMRI studies suggesting the highest multisensory integration in terms of neural activation for congruent mouth-voice stimuli (e.g.,[14]).

In contrast to previous studies showing that the integration window breaks down for distances greater than 20-25 degrees [4], in our case the magnitude of the visual bias is significant also when the incongruent AV stimuli are coming from the two avatars at the extremes of the screen. This can be interpreted in terms of synchronized AV pairs being

merged as a single event irrespective of their spatial disparity due to their temporal correlation perceived as a form of causation [8]. We can argue that embedded semantics in the scene contribute to a wider integration window with respect to the boundaries empirically found for simplified stimuli [1]. To further verify this hypothesis, we examined whether the induced bias during incongruent AV stimuli was in the direction of the visual cue, i.e., we tested the ventriloquist effect. The results for the different conditions are shown in Fig. 3.d, which are consistent with the hypothesis that visual cues encoding environmental statistics induce a stronger bias and that the magnitude of the bias is related to the embedded semantics, e.g., *moving lips+arm* induces a slightly stronger bias than *moving lips*. Furthermore, while an incongruent arm movement in *moving lips–arm* acts as distractor decreasing the magnitude of the visual bias towards the lips, this magnitude is still significant.

Finally, when the auditory cue is played along with static avatars, subjects were not as accurate as expected in the absence of a visual bias. One hypothesis for this effect is that the extended exposure to animated avatars may create the expectation of seeing similar animated patterns in the next trials, thus perceiving a static avatar as incongruent with respect to an expected dynamic visual cue.

3. Neurorobotic Experiment

The goal of the neurorobotic experiment was to trigger human-like responses with an iCub [5] exposed to the same conditions as the human subjects. We used the collected behavioural data to train a deep learning model and compared the results with human responses. We propose a deep learning model processing both spatial and feature-based information in which low-level areas (such as the visual and auditory cortices) are predominantly unisensory, while neurons in higher-order areas encode multisensory representations. The proposed architecture comprises 3 input channels (audio, face, and body motion) and a hidden layer that computes a discrete behavioural response on the basis of the output of these unisensory channels (see Fig. 2).

To train our model on crossmodal conflict resolution, we first train the individual channels using modality-specific spatial information (Fig. 2; gray bounding box). The auditory channel is trained to locate a sound source, the face channel to locate moving lips, and the body motion channel to locate arm movement. This procedure ensures that each channel is able to describe modality-specific stimuli. After the individual training of these 3 channels, a fully connected hidden layer receives modality-specific representations as input and is trained using the human responses as teaching signals. The output softmax layer represents a probability distribution over the 4 possible responses.

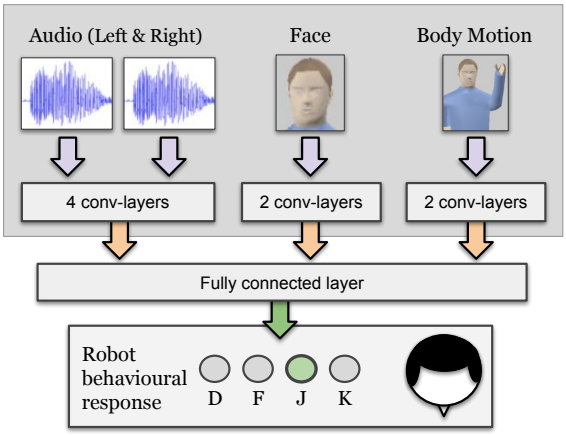

Figure 2: Multichannel deep learning model for multisensory integration and conflict resolution. The model combines sound localization, facial features, and body motion to produce a discrete behavioural response in real time. Each channel is first trained with modality-specific spatial information (gray bounding box) and used as input for a hidden layer trained with multisensory representations using human behavioural responses as the teaching signals.

3.1. Robot Behaviour

For a direct human-robot comparison, we placed the iCub in front of the projection screen (Fig. 3.a; see Fig. 1.a for setup with humans). In order to prevent biasing the robot behaviour towards a specific subject, we evaluate the model using *leave-one-out* cross-validation with the responses of 32 subjects for training and of 1 subject for testing. Since each participant produced 600 responses, we had 32×600 training data points for each training fold, for which a new network was initialized. This training procedure resulted in 33 network instances from which we generated 33×600 responses used to compare robot-vs-human behaviour.

The error rates of the robot averaged across all conditions are shown in Fig. 3.b, where it can be seen that the human-vs-robot ER difference is not significant. Interestingly, there is an inverse trend with respect to humans in which the ER is higher for the avatars in the center and decreases for the ones at the sides. This difference can be explained due to the different ways in which humans and the robot process incoming visual input. Human vision has higher spatial resolution towards the center of the FOV (referred to as foveal vision), which leads to a stronger visual bias over the estimate of the sound source's location when the visual cue occurs towards the center [7]. On the other hand, the visual input processed by the robot does not comprise such foveal property, and consequently, the visual bias has the same magnitude irrespective of its position within the FOV of the camera. However, since the model is trained with data collected from robot sensors but using human responses as a

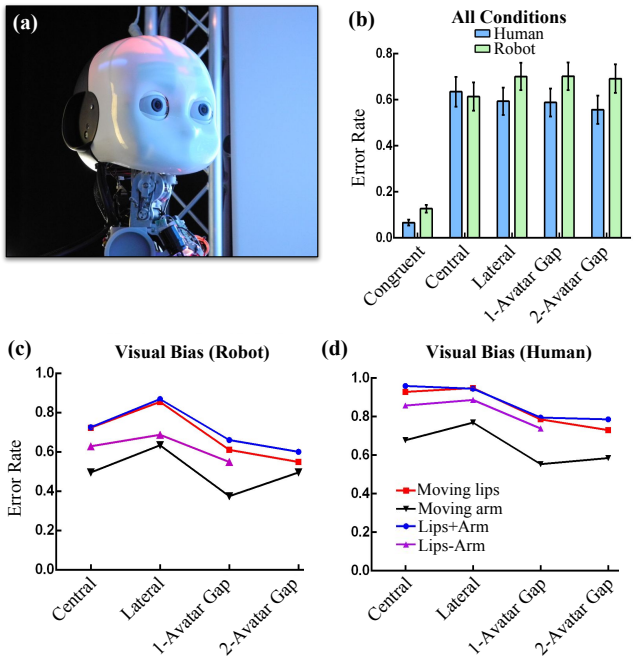

Figure 3: Neurorobotic experiment: (a) iCub exposed to the immersive experimental setup (b) Error rate after exposure to congruent and incongruent AV stimuli (c) Robot vs Human comparison in terms of the ventriloquism effect (response biased towards the visual cue) showing similar trends for all the conditions with incongruent AV stimuli.

teaching signal, there is a compensation artefact introduced by the hidden layer which results in such an inverse trend of the magnitude of the visual bias in relation to its position. In order to address this artefact, it would be necessary to model properties of foveal vision embbeded in the convolutional channels processing the visual input (e.g., [3]).

In terms of the magnitude of the bias reflecting environmental statistics, we analyzed the proportion of ER due to shifting the estimate towards the visual cue (ventriloquism effect). It can be seen from Fig. 3.c-d that the behaviour of the robot resembles human responses for all the conditions.

4. Future Work

The obtained results motivate further research in three main directions. First, the experimental scenario can be extended to more natural scenes, e.g., by displaying real-world videos with human characters. Second, the deep learning model was trained in a supervised fashion, i.e., by providing the expected responses as a target. Instead, it may be of interest to study whether and how such behaviour can emerge from the unsupervised exposure to congruent AV stimuli, e.g., by learning environmental statistics. Third, we observed that subjects adopting a strategy that relied mostly on auditory cues exhibited smaller error rates. Studies suggest that the brain changes its strategy according to the reliability of sensory drive and mechanisms of cognitive control [13]. Consequently, it would be of interest to model the dynamic selection of perceptual strategies on the basis of modality-specific reliability, conflict adaptation effects, top-down attention, and prior knowledge.

References

[1] C. Dahl, N. Logothetis, and C. Kayser. Modulation of visual responses in the superior temporal sulcus by audio-visual congruency. *Front Integr Neurosci.*, 4:10, 2010. 1, 3

[2] K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams. Causal inference in multisensory perception. *PLOS ONE*, 2(9):1–10, 2007. 1

[3] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel. Understanding the effective receptive field in deep convolutional neural networks. *CoRR*, abs/1701.04128, 2017. 4

[4] E. Magosso, C. Cuppini, and M. Ursino. A neural network model of ventriloquism effect and aftereffect. *PLOS ONE*, 7(8):1–19, 2012. 2

[5] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano. The icub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks*, 23(8):1125 – 1134, 2010. 3

[6] K. Noda, H. Arie, Y. Suga, and T. Ogata. Multimodal integration learning of robot behavior using deep neural networks. *Robot Auton Syst.*, 62(6):721 – 736, 2014. 1

[7] B. Odegaard, D. Wozny, and L. Shams. Biases in visual, auditory, and audiovisual perception of space. *PLoS Computational Biology*, 11(12), 2015. 2, 3

[8] C. V. Parise, C. Spence, and M. O. Ernst. When correlation implies causation in multisensory integration. *Current Biology*, 22(1):46 – 49, 2012. 3

[9] G. Parisi, P. Barros, D. Fu, S. Magg., H. Wu, X. Liu, and S. Wermter. A neurorobotic experiment for crossmodal conflict resolution in complex environments. IEEE/RSJ IROS, Madrid, Spain., 2018. 1

[10] G. I. Parisi, P. Barros, M. Kerzel, H. Wu, G. Yang, Z. Li, X. Liu, and S. Wermter. A computational model of crossmodal processing for conflict resolution. In *IEEE EPIROB-ICDL*, pages 33–38. IEEE, 2017. 1

[11] D. B. Polley. Multisensory conflict resolution: Should I stay or should I go? *Neuron*, 93(4):725 – 727, 2017. 1

[12] M. Vavrečka and I. Farkaš. A multimodal connectionist architecture for unsupervised grounding of spatial language. *Cognitive Computation*, 6:101–112, 2013. 1

[13] G. Yang, W. Nan, Y. Zheng, H. Wu, Q. Li, and X. Liu. Distinct cognitive control mechanisms as revealed by modality-specific conflict adaptation effects. *Journal of Experimental Psychology: Human Perception and Performance*, 43(4):807–818, 2012. 4

[14] L. L. Zhu and M. S. Beauchamp. Mouth and voice: A relationship between visual and auditory preference in the human superior temporal sulcus. *Journal of Neuroscience*, 2017. 1, 2