International Journal of Social Robotics (2023) 15:1325–1340
https://doi.org/10.1007/s12369-023-00993-3

A Trained Humanoid Robot can Perform Human-Like Crossmodal Social Attention and Conflict Resolution

Di Fu[1,2,3] · Fares Abawi[3] · Hugo Carneiro[3] · Matthias Kerzel[3] · Ziwei Chen[1,2] · Erik Strahl[3] · Xun Liu[1,2] · Stefan Wermter[3]

Accepted: 8 March 2023 / Published online: 2 April 2023

Abstract

To enhance human-robot social interaction, it is essential for robots to process multiple social cues in a complex real-world environment. However, incongruency of input information across modalities is inevitable and could be challenging for robots to process. To tackle this challenge, our study adopted the neurorobotic paradigm of crossmodal conflict resolution to make a robot express human-like social attention. A behavioural experiment was conducted on 37 participants for the human study. We designed a round-table meeting scenario with three animated avatars to improve ecological validity. Each avatar wore a medical mask to obscure the facial cues of the nose, mouth, and jaw. The central avatar shifted its eye gaze while the peripheral avatars generated sound. Gaze direction and sound locations were either spatially congruent or incongruent. We observed that the central avatar's dynamic gaze could trigger crossmodal social attention responses. In particular, human performance was better under the congruent audio-visual condition than the incongruent condition. Our saliency prediction model was trained to detect social cues, predict audio-visual saliency, and attend selectively for the robot study. After mounting the trained model on the iCub, the robot was exposed to laboratory conditions similar to the human experiment. While the human performance was overall superior, our trained model demonstrated that it could replicate attention responses similar to humans.

Keywords Crossmodal social attention · Eye gaze · Conflict processing · Saliency prediction model · iCub robot

1 Introduction

Robots are increasingly becoming an integral part of daily life. It is essential for robots to behave as social actors capable of processing multimodal social cues, enriching interactions with humans. Moreover, to understand humans' intentions, it is crucial to explore how they process information and the underlying cognitive mechanisms behind it [46]. The need for such solutions encourages the design of socially functional robots to meet more significant challenges and difficulties in human-robot communication.

The current study adopts a dynamic variant of the gaze-triggered Posner cueing paradigm [53] for testing the attentional orienting effect of eye gaze on auditory target detection. We construct a synthetic scenario using the framework introduced by Parisi et al. [50] and Fu et al. [23] to study crossmodal spatial attention for sound localisation. In the aforementioned studies, a 4-avatar round-table meeting

✉ Di Fu
di.fu@uni-hamburg.de

✉ Xun Liu
liux@psych.ac.cn

Fares Abawi
fares.abawi@uni-hamburg.de

Hugo Carneiro
hugo.carneiro@uni-hamburg.de

Matthias Kerzel
matthias.kerzel@uni-hamburg.de

Ziwei Chen
czwscda2015@163.com

Erik Strahl
erik.strahl@uni-hamburg.de

Stefan Wermter
stefan.wermter@uni-hamburg.de

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Beijing, China

[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing, China

[3] Department of Informatics, University of Hamburg, Hamburg, Germany

scenario experiment was conducted on human participants. During the task, lip and arm movements were used as the visual cues, either spatially congruent or incongruent with the auditory target. Our previous findings indicated that lip movement was more salient than arm movement, implying a stronger visual bias on auditory target localisation. This is due to the physical association between lip movement and speech [77]. Furthermore, previous research also revealed that head orientation was a primary social cue for triggering the reflexive attention of an observer [38]. To align our experimental setup with the Posner gaze-cueing task, we reduce the number of avatars to three. The central avatar shifts its eyes with a slight tilt in its head and upper body posture towards the direction of gaze. To avoid distractions from lip movements, all three avatars wear medical masks to obscure their faces partially. The current social norm inspires this task design. In multiperson social contexts during the COVID-19 pandemic, the use of medical masks is common. Research shows that wearing masks decreases both adults' and children's face recognition abilities [26, 68]. As a result, humans have to rely on gaze cues to compensate for the lack of lip movement in identifying social intentions [16].

For the robotic experiment in this work, an iCub head is used to emulate human social attention [49]. We modify the Gated Attention for Saliency Prediction (GASP) model [1] and mount it on the iCub head to predict crossmodal saliency. GASP can detect multiple social cues, producing feature maps for each. These maps are prioritised based on a weighting mechanism to mitigate stronger cues. Following the weighting stage, the features are sequentially integrated, and the model is trained on eye tracking data to predict saliency. The iCub gaze movements are based on the saliency density maps predicted by the GASP model.

We define two goals for our current study. First, we aim to detect human responses for a crossmodal social attention task with dynamic stimuli to determine the eye gaze orienting effect on sound localisation. Second, we emulate human behavioural patterns using a humanoid robot, running a social attention model which is tested in similar laboratory conditions. Thus, human and robot responses are compared under congruent and incongruent audio-visual localisation conditions in the gaze-cueing task. In this study, the Stimulus-Response Compatibility (SRC) effect [54] is measured to detect the conflict resolution ability of the participants and the iCub robot. This effect occurs when stimulus and response in an SRC paradigm are spatially incongruent. Participants show poorer performance (e.g., lower accuracy and slower response to stimuli) under incongruent conditions compared with congruent conditions [4]. Larger SRC effects indicate weaker conflict resolution ability [40]. Previous research also set a neutral condition as a baseline to distinguish whether irrelevant or incongruent stimuli cause the SRC effect entirely [60]. If there is no significant difference between participants' performance under neutral and congruent conditions, the SRC effect comes from irrelevant or incongruent stimuli interference. If the performance of the neutral condition is significantly worse than the congruent condition, it means that congruent stimuli have a facilitation effect on conflict processing [37, 41]. Thus, we set a neutral condition to study whether there is an interference or facilitation effect where the central avatar does not shift its eyes, head, or upper body in any direction in the current study.

According to our research goals, the current study proposes the following hypotheses:

In the human experiment:

H1: Eye gaze can trigger the attentional orienting effect, which leads to better performances with the congruence of gaze direction and auditory targets.

H2: For the neutral condition, no irrelevant visual stimulus shows up before the auditory target. We assume that participants' performance in the neutral condition might be intermediate between congruent and incongruent conditions. More specifically, no significant difference between performance under congruent and neutral conditions, suggests that the SRC effect is from the interference of the incongruent condition.

In the robot experiment:

H3: Modelling the reflexive attentional orienting effect is achievable by integrating a binaurally aware auditory localiser for estimating the direction of sound arrival.

H4: A neurocognitive model trained on human eye fixations can result in a robot attentional orientation consistent with human responses under the congruent, incongruent, and neutral conditions.

To test the validity of these hypotheses, the article is structured into two parts. The first part focuses on how humans behave in a crossmodal conflict task triggered by eye gaze as a visual cue. Background on the use of eye gaze as a social cue is provided in Sect. 2. The full description of the experiment performed with human participants and the results achieved are provided in Sect. 3. The following part of the article focuses on whether a robot can behave similarly to a human in the same experimental scenario. For that, a description of GASP, the attention mechanism used by the robot, is presented in Sect. 4, and the setup of the robotic experiment, as well as a comparison between the performances of the robot and those of the human participants, are presented in Sect. 4.3.2. Finally, Sect. 5 offers a discussion on the achieved results, and Sect. 6 indicates potential future research directions.

2 Background and Related Work

2.1 Social Attention

Social attention is the ability to follow others' eye gaze and infer where and what they are looking at [10]. Social attention is the fundamental function of sharing and conveying information with other agents, contributing to the functional development of social cognition [44]. Social attention allows humans to quickly capture and analyse others' facial expressions, voices, gestures, and social behaviour, so that they can participate in social interaction and adapt within society [38, 39]. Furthermore, this social function enables the recognition of others' intentions and the capture of relevant occurrences in the environment (e.g., frightening stimuli, novel stimuli, and reward) [49]. The neural substrates underlying social attention are brain regions responsible for processing social cues and encoding human social behaviour, including the orbital frontal and middle frontal gyrus, superior temporal gyrus, temporal horn, amygdala, anterior precuneus lobe, temporoparietal junction, anterior cingulate cortex, and insula [3, 49]. From a developmental perspective, infants' attention to social cues helps them quickly learn how to interact with others, learn a language, and build social relationships [66]. However, dysfunctional social attention is one of the primary social impairments for children with Autism Spectrum Disorder (ASD) [67]. For example, infants with (ASD) are born with less attention to social cues, an inability to track the sight of others, and a fear of looking directly at human faces [61]. This might be a crucial mechanism that results in their failure to understand others' intentions and engage in typical social interactions [67]. Research on developmental mechanisms of social attention is still in its early stages. Exploring these scientific questions will be significant for understanding mechanisms of interpersonal social behaviour and developing clinical interventions to assist individuals diagnosed with ASD.

2.2 Eye Gaze as Social Cue

One of the most critical manifestations of social attention is the ability to follow others' eye gazes and respond accordingly [62]. Eye gaze is proven to have higher social saliency and prioritisation than other social cues [38] since it indicates to a person the direction in which another person is looking [22]. Gaze following is considered as the foundation of more sophisticated social and cognitive functions like the theory of mind, social interaction, and survival strategies formed by evolution [7, 38]. For instance, infants can track the eye gaze of their parents at the age of 3 months [19, 32, 33]. After 10 months, gaze following ability significantly contributes to their language development [11, 62]. Psychological studies use the modified Posner cueing task [52] or named gaze-cueing task [20] to study reflexive attentional orienting generated by the eye gaze. During the task, the eye gaze is presented as the visual cue in the middle of the screen, followed by a peripheral target, which could be spatially congruent (e.g., a right-shift eye gaze followed by a square frame or a Gabor patch shown on the right side of the screen), or incongruent. However, studying the visual modality alone is not enough to reveal how humans can quickly recognise social and emotional information conveyed by others in an environment full of multimodal information [8]. Selecting information from the environment across different sensory modalities allows humans to detect crucial information such as life threats, survival strategies, etc. [24, 45]. Therefore, several studies conducting a crossmodal gaze-cueing task demonstrate the reflexive attentional effect of the visual cue on the auditory target [17, 42]. Most of these studies rely on images of gaze shifts as visual cues to trigger the observers' social attention [45, 48]. However, these images are not dynamic and lack ecological validity.

2.3 Stimulus–Response-Compatibility tasks and Effects

Researchers study humans' cognitive control mechanism by using the Stimulus–Response-Compatibility (SRC) tasks to measure the behavioral performance and neural activation on conflict processing. The SRC effect measured by those tasks reflects humans' better performance in the Stimulus–Response congruent conditions than the incongruent conditions. The classic SRC tasks conducted in the lab are Stroop task [69], Flanker task [18], and Simon task [64]. The size of the SRC effect represents the capacity of conflict processing. The larger SRC effect may be accompanied with the weaker top-down control, dysfunction or immaturity of conflict control [14, 43].

2.4 Audio-Visual Saliency Modelling

Saliency prediction models are trained on eye tracking data collected from multiple participants looking at images or videos under the free-viewing condition. Several studies show that audio-visual input improves models' performances in predicting saliency. Tavakoli et al. [70] propose a late fusion audio-visual model for enhancing saliency prediction compared to visual-only models. Tsiami et al. [71] show that the early fusion of auditory and visual stimuli reduces reliance on visual content when inferring salient regions. Jain et al. [31] compare multiple approaches for integrating the two modalities within different layers of the model hierarchies. In contrast to previous findings, the authors show that auditory input degrades performance, suggesting that better audio-visual integration methods are needed. Moreover, sound localisation performances of monaural audio-visual models cannot surpass binaural audio-visual models [55, 75]. This is due to the reduced ability of monaural models to accurately localise sound since the interaural temporal and level difference cannot be computed [73]. Since our task relies mainly on sound direction, we design a binaural sound localisation model that infers saliency both from auditory and visual stimuli.

3 Human Experiment

3.1 Participants

37 participants (female = 20) participated in this experiment. Participants were between 18 to 29 years of age, with a mean age of 22.89 years. All participants reported no history of neurological conditions (seizures, epilepsy, stroke, etc.) and had either normal or corrected-to-normal vision and hearing. This study was conducted following the principles expressed in the Declaration of Helsinki. Each participant signed a consent form approved by the Ethics Committee of the Institute of Psychology, Chinese Academy of Sciences.

3.2 Experimental Setup

All participants watch clips under normal indoor light conditions. Auditory noise in their surroundings is minimal, and the room acoustic effects are negligible since the sound is played directly through on-ear headphones. This section describes the stimuli generation procedure, the environmental setup, and the data recording methodology.

3.2.1 Apparatus, Stimuli and Procedure

Virtual avatars are chosen over recordings of real people, as the experiment requires strict control over the avatar's behaviour, both in terms of timing and exact motion. By using synthetic data as the experimental stimuli, it can be ensured, for instance, that looking to the left and right are exactly symmetrical motions, thus avoiding any possible bias. Moreover, using three identical avatars that are only different in terms of clothing colour also alleviates a bias towards individuals in a real setting. The static basis for the highly-realistic virtual avatars was created in MakeHuman.[1] Based on these avatar models, a data generation framework for research on shared perception and social cue learning with virtual avatars [34] (realised in Blender[2] and Python) is used to create the animated scenes with the avatars, which are used as the experimental stimuli in this study. The localised sounds are created from a single sound file using a head-related transfer function[3] that modifies the left and right audio channels to simulate different latencies and damping effects for sounds arriving from different directions. In our 3-avatar scenario, the directions are frontal left and frontal right at 60 degrees, corresponding to the positions where the peripheral avatars stand.

During the experiment, the participants sit positioned 55 cm from the monitor at a desk and wear headphones, as depicted in Fig. 1a. In each trial, a fixation cross appears in the middle of the screen for 100–300 ms with equal probability. Next, a visual cue is displayed for 400 ms, consisting of an eye gaze shift and a synchronised slight head and upper body shift from the central avatar. In each trial, the central avatar randomly chooses to look at the avatar at the right, at the one at the left, or directly towards the participant, meaning no eye gaze shift at all. Afterwards, the left or the right avatar says "hello" with a human male voice as the auditory target. This step lasts for 700 ms. Finally, another fixation cross is shown at the centre of the screen for 700, 800 or 900 ms, with equal probability, until the end of the trial (cf. Fig. 1c for a schematic representation of the trial).

The experimental design has three directions for the visual cue (left, right, and central) and two for the auditory target location (left, right). The congruent audio-visual condition occurs when the central avatar's eye gazes in the same direction as the avatar who generates the sound. The incongruent audio-visual condition occurs when the central avatar's eye gazes in the opposite direction as the avatar who generates the sound. The neutral condition is when the central avatar does not shift its eye gaze, so there is no spatial conflict between the visual cue and the following auditory target. The participants begin the experiment with 30 practice trials and enter into the formal test when their accuracy of practice trials reaches 90%. Each condition is repeated 96 times, with a total of 288 trials separated into four blocks. There is a 1-minute rest between every two blocks. The time duration for each trial is 1900–2300 ms, and the formal test lasts for 12 min.

During the task, the participants are asked to determine as soon and precisely as possible whether the auditory stimulus originated from the avatar on the left or on the right. The participants make decisions by pressing the keys "F" and "J" on the keyboard, corresponding to the left and right avatars. The participants' responses during the display of the auditory target and the second fixation are recorded. The stimulus display and response recording are both under the control of

[1] http://www.makehumancommunity.org/

[2] https://www.blender.org/

[3] https://sound.media.mit.edu/resources/KEMAR.html

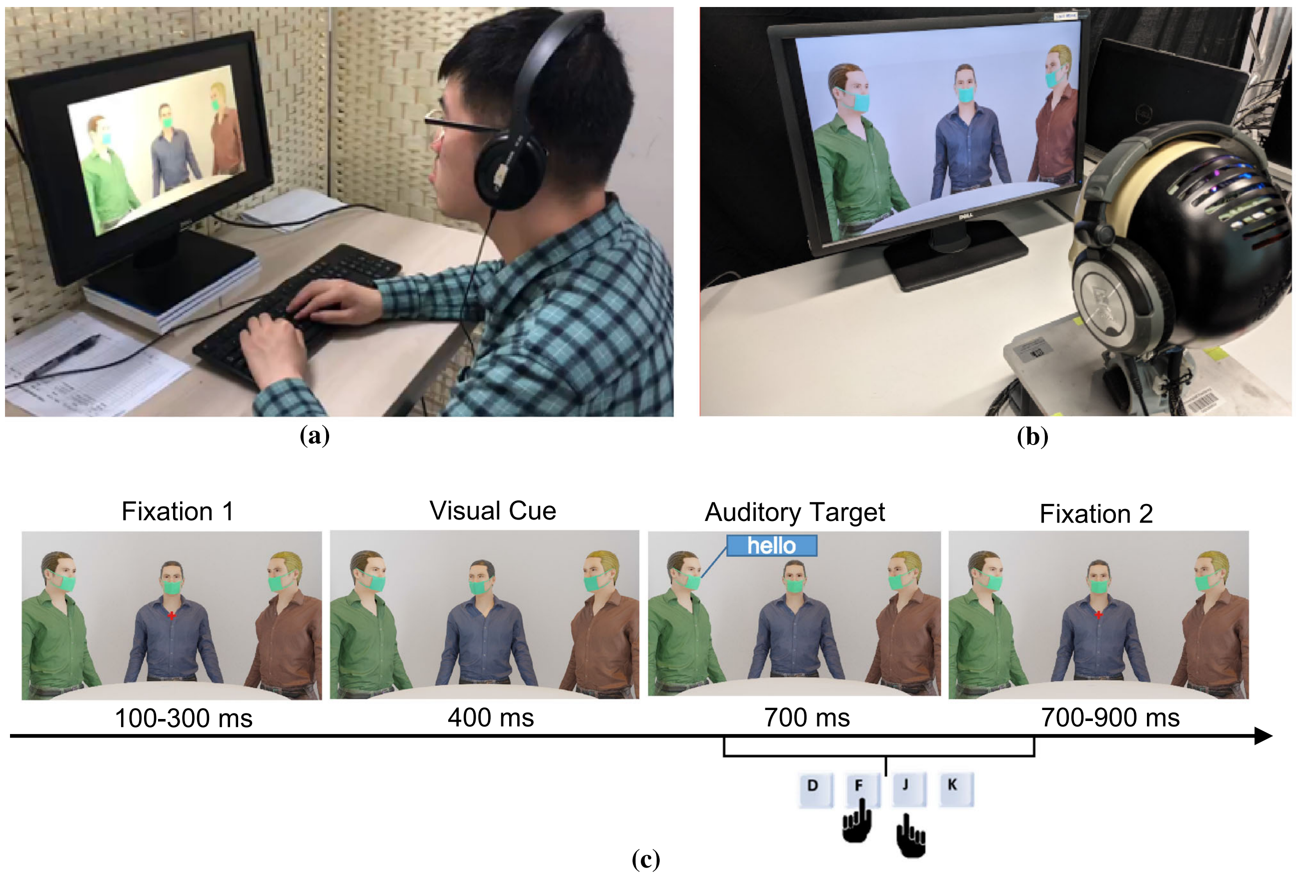

Fig. 1 Audio-visual gaze-cueing social attention task. **a** Participant engaging in the formal test with a headphone to hear the auditory stimulus, and a keyboard to respond; **b** The iCub robot engaging in the test with a headphone to get the auditory input, and responding to the target by moving its eyeballs (see Fig. 4); **c** Schematic illustration of a single trial

E-prime 2.0.[4] In the current study, all participants perceive the simulated masks as typical.

3.2.2 Data Recording and Analyses

Reaction time (RT) and error rates (ER) are analysed as human response indices. For the RT analysis, error trials, and trials with RTs shorter than 200 ms, and those with RTs beyond three standard deviations above or below the mean were excluded, corresponding to 2.42% of the data being removed. To examine the Stimulus–Response Compatibility effects of the crossmodal audio-visual conflict task, one-way repeated measures analysis of variance (ANOVA) is used to test differences in the participants' responses under the three congruency conditions (congruent, incongruent and neutral). All post hoc tests in the current study use Bonferroni correction.

3.3 Experimental Results

Our experimental results indicate that the participant response time and accuracy under the audio-visually congruent condition exceeded the performance under the incongruent condition. There are no significant differences between the neutral and incongruent conditions for both RT and ER. The lack of difference between the neutral and incongruent conditions shows that the lack of congruent audio-visual cueing negatively affects the participants' performance.

3.3.1 Reaction Time

A repeated measures ANOVA with a Greenhouse-Geisser correction shows that the participants' RT differs significantly between different congruency conditions, $F\ (2, 34) = 24.19$, $p\ <\ .001$, $\eta_p^2\ =\ .40$ (see Figs. 2a and b). Post hoc tests show that the participants responded significantly faster under the congruent condition (mean $\pm$ SE $=$ 466.25 $\pm$ 14.92 ms) than both incongruent condition (mean $\pm$ SE $=$ 485.12$\pm$14.82 ms, $p\ <\ .001$) and neutral condition (mean $\pm$

[4] Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime (Version 2.0). [Computer software and manual]. Pittsburgh, PA: Psychology Software Tools Inc.

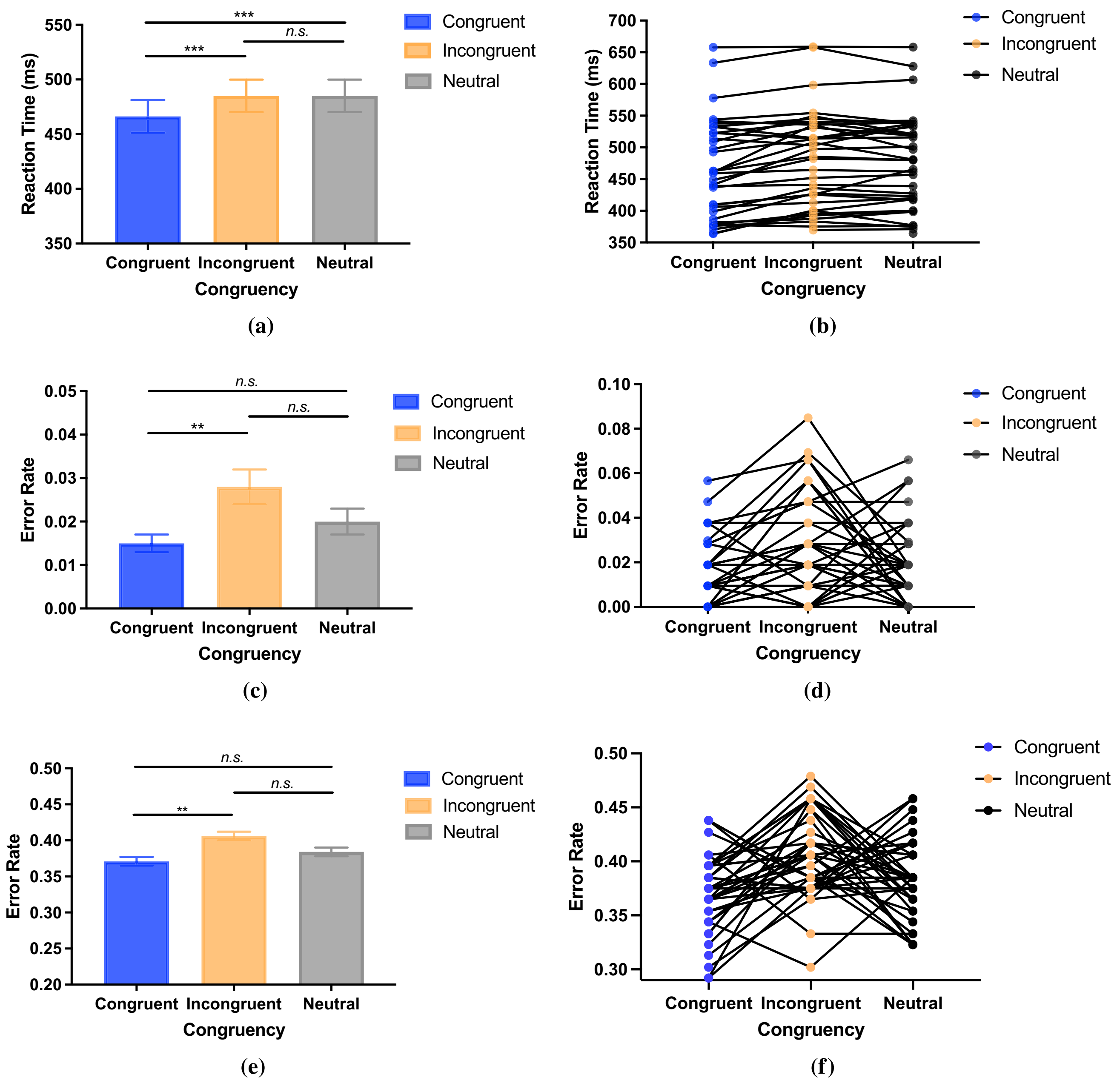

Fig. 2 **a** RT of participants under different congruency conditions – group level; **b** RT of participants under different congruency conditions – individual level; **c** ER of participants under different congruency conditions – group level; **d** ER of participants under different congruency conditions – individual level; **e** ER of the iCub under different congruency conditions – group level; **f** ER of the iCub under different congruency conditions – individual level. $*$ denotes $.01 < p < .05$, $**$ $.001 < p < .01$, $***$ $p < .001$, and *n.s.* denotes no significance

SE $= 485.11 \pm 14.80$ ms, $p < .001$). However, the difference between the incongruent and neutral condition was not significant, $p > .05$.

3.3.2 Error Rates

A repeated measures ANOVA with a Greenhouse-Geisser correction shows that the participants' ER differs significantly between different congruency conditions, $F\,(2, 34) = 5.69$, $p < .05$, $\eta_p^2 = .14$ (see Fig. 2c and d). Post hoc tests show that the participants presented significantly lower ER under the congruent condition (mean $\pm$ SE $= .02 \pm .002$) than the incongruent condition (mean $\pm$ SE $= .03 \pm .004$), $p < .01$. However, there was no statistical significance in the difference between the neutral condition (mean $\pm$ SE $= .02 \pm .003$) and both other congruency conditions, $p > .05$ in both cases.

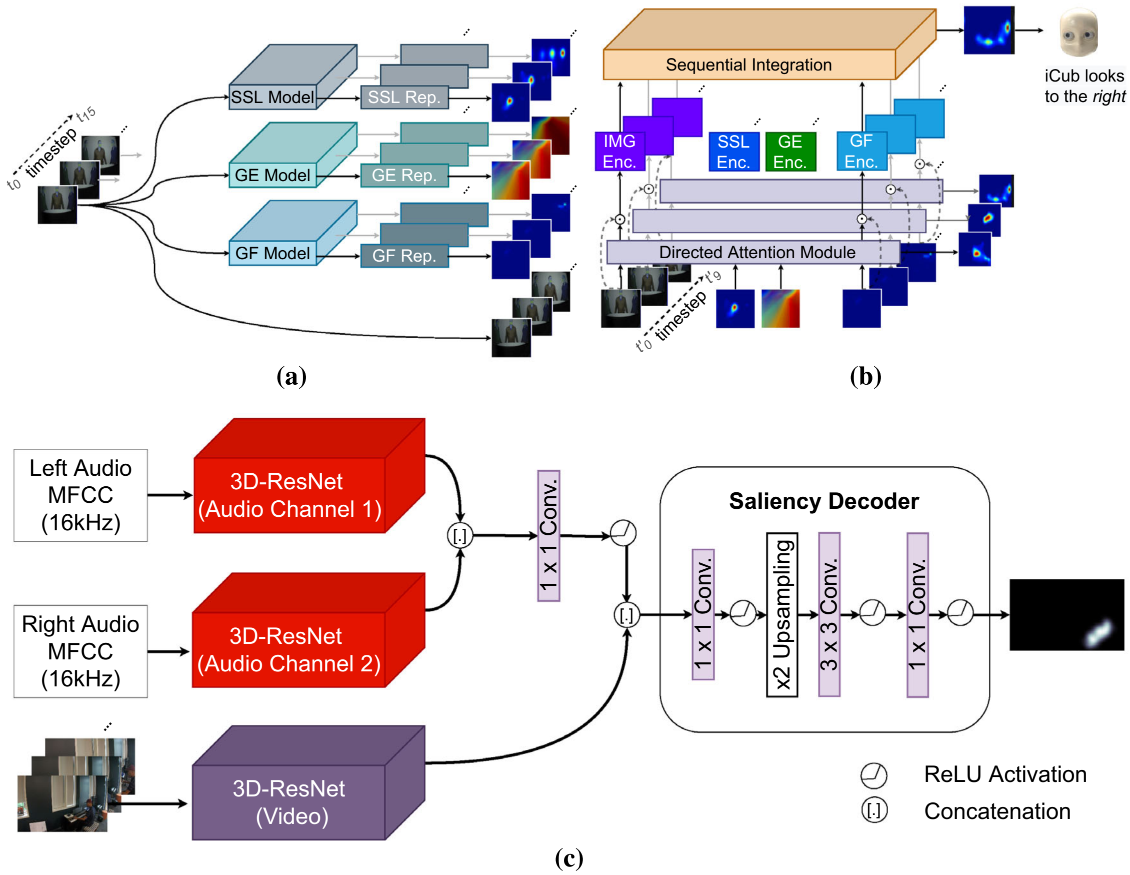

Fig. 3 **a** SCD – Social cue detection stage in which the representations of the sound source localisation (SSL), gaze estimation (GE), and gaze following (GF) are extracted; **b** GASP – Saliency prediction; **c** Binaural DAVE – Audio-visual sound source localisation

4 Robot Experiment

4.1 Neural Modelling

To assess whether the iCub head would display degradation in performance under the incongruent condition relative to the congruent condition, a model capable of dealing with stimuli from the gaze following modality showing the attention targets of all individuals observed in the video, as well as the gaze estimation modality, indicating their head and eye poses as well as audio source localisation, was needed. For that sake, we opted for using the GASP model [1], which showed a high performance when dealing with gaze and audio-visual stimuli. However, GASP was originally projected to work solely with monaural inputs. Since the auditory stimulus in the three-avatar scenario arrives from a single direction, we modify GASP to accommodate stereo audio. We do so by replacing the saliency prediction model with a binaural sound localisation model.

4.1.1 Dynamic Saliency Prediction

The process of predicting saliency is divided into two stages. The first stage, Social Cue Detection (SCD), is responsible for extracting social cue feature maps from a given audio-visual sequence. Figure 3a depicts the architecture of the SCD stage. Given a sequence of images and their corresponding high-level feature maps, the second stage, GASP, then predicts the corresponding saliency region by integrating the social cue feature map sequences. The overall integration pipeline followed by GASP is shown in Fig. 3b.

Following the implementation of GASP, the SCD stage comprises four modules, each responsible for extracting a specific social cue [1]. Those modules include gaze following, gaze estimation, facial expression recognition, and audio-visual saliency prediction. For the current task, however, the facial expression recognition module is not employed since the virtual avatar faces are partially occluded and do not display facial expressions. In order to closely replicate the experiments done with participants, the iCub robot receives auditory stimuli from both ears. An audio-visual saliency prediction module was originally designed to work with monaural stimuli. To operate on binaural stimuli, we replace the saliency prediction module with a binaural audio-visual sound source localisation (SSL) model, denoted the "SSL model" in Fig. 3a. The binaural SSL model architecture is shown in Fig. 3c.

The video streams used as input are split into their frames and corresponding auditory signals. For every video frame and corresponding audio signal, the SCD stage covers the extraction of social cue feature maps, which are then propagated to GASP. The Directed Attention Module (DAM) weighs the feature map channels to emphasise those that represent high unexpectedness with respect to their predictions. Convolutional layers further encode those weighted feature map channels. In Fig. 3b, these layers are denoted by "Enc." (for encoder). The encoded feature maps of all video frames are then integrated using a recurrent extension of the convolutional Gated Multimodal Unit (GMU) [6]. The GMU's mechanism weighs the features of its inputs. Adding a convolutional aspect to it accounts for the preservation of spatial properties of the input features. The recurrent property of the integration unit considers the whole sequence of frames by performing the gated integration at every timestep.

For this work, the LARGMU (Late Attentive Recurrent Gated Multimodal Unit) is used because of its high performance compared to other GMU-based models [1]. Since LARGMU is based on the convolution GMU, it preserves the input spatial features. The LARGMU's recurrent structure allows it to integrate those features sequentially. Adding a soft-attention mechanism based on the convolutional Attentive Long Short-Term Memory (ALSTM) [15] prevents gradients from vanishing as feature sequences get sufficiently large. As the name implies, LARGMU is a late fusion unit, meaning that the gated integration is performed after the input channels are concatenated and, in sequence, propagated to the ALSTM.

4.1.2 Binaural Sound Localisation

DAVE (Deep Audio-Visual Embedding) [70] is used as a sound source localisation module in the SCD stage. In its original form, the audio-visual DAVE encodes inputs from one video and one audio stream, which are projected into a feature space by 3D-ResNets [28] (one for each input stream). 3D-ResNet extends the ResNet model [30] to operate on multiple frames by replacing 2D convolutional layers with their 3D counterparts. Its encoder is followed by a convolutional saliency decoder that upscales the latent representation and provides the corresponding saliency map. For our current work, DAVE is extended to accept binaural input, and this binaural extension structure uses a similar rationale to the monaural DAVE, see Fig. 3c. The main difference is using two 3D-ResNets to process the auditory modality, whose output features are concatenated and then encoded and downsampled by a two-dimensional convolutional layer. This layer is responsible for guaranteeing that the dimension of the feature produced by this part of the architecture matches that of the feature produced by DAVE's original audio-stream 3D-ResNet.

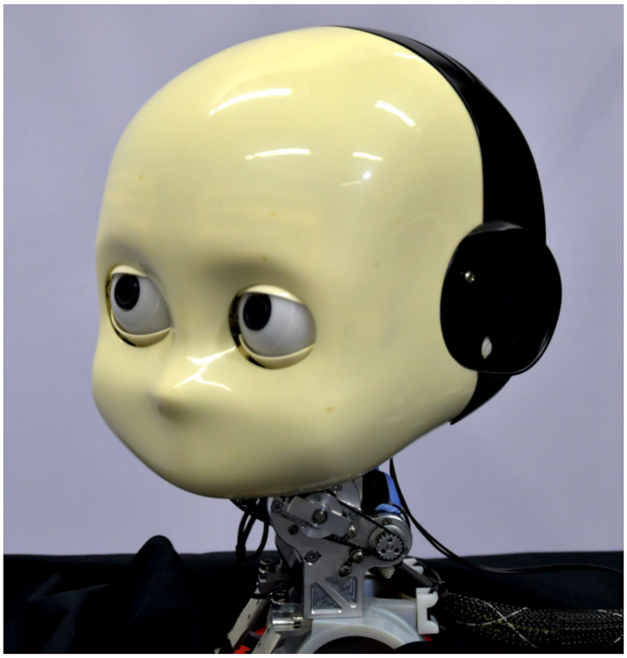

Fig. 4 The iCub head

We initialise the binaural DAVE with the pre-trained parameters of the audio-visual DAVE [70]. The left and right auditory streams are initialised with identical parameter weights extracted from the 3D-ResNet auditory stream of the monaural variant. The 1 × 1 convolutional layer that encodes the concatenated audio features is initialised using the normalisation method proposed by He et al. [29]. All model parameters are optimised except for the video 3D-ResNet, which are frozen throughout optimisation following DAVE's training procedure [70].

4.1.3 Binaural DAVE as a Prior to GASP

The GASP architecture used in the experimental setup consists of the pretrained GASP, excluding the facial expression recognition input stream. We replace the audio-visual saliency detector with DAVE's binaural sound localisation variant. Abawi et al. [1] show that replacing saliency predictors does not require re-training GASP, allowing us to use a sound localisation model in the place of a saliency prediction model without fine-tuning the sequential integration parameters.

GASP receives four sequences of data as input, one sequence of consecutive frames of the original video and three sequences of feature maps, one for each model in the social cue detection stage. In our experiment, we capture sequences of 10 frames (cf. timesteps t'_0 to t'_9 in Fig. 3b). The number of frames received as input by each model in the SCD stage varies due to dissimilarity in their expected inputs. The sound localisation model receives a sequence of 16 frames as input, whereas the gaze estimation and following models receive sequences of 7 frames each. A more detailed explanation of how the frames are selected based on the timestep being processed is provided by Abawi et al. [1]. The auditory input is captured as a one-second chunk and propagated to each audio 3D-ResNet of the sound localisation model. In this experiment, GASP is embedded in the iCub robot and subjected to the same series of one-second videos as the participants. The one-second chunk used as input to the binaural sound localisation model corresponds to the entire audio recording per video.

4.2 iCub Eye Movement Determination

After the iCub acquires the visual and auditory inputs, the social cue detectors and the sound source localisation model extract features from those audio-visual frames. Following the detection and generation of the feature maps, they are propagated to GASP, which, in turn, predicts a fixation density map $\mathcal{F}\colon \mathbb{Z}^2 \to [0, 1]$, which is displayed in the form of a saliency map for a given frame. The fixation peak $(x_\mathcal{F}, y_\mathcal{F})$ is determined by calculating

$$(x_\mathcal{F}, y_\mathcal{F}) = \text{argmax}_{x,y}\, \mathcal{F}(x, y)\,. \tag{1}$$

The values of $x_\mathcal{F}$ and $y_\mathcal{F}$, originally in pixels, are then normalised to scalar values $\hat{x}_\mathcal{F}$ and $\hat{y}_\mathcal{F}$ within the $[-1, 1]$ range, such that

$$\hat{x}_\mathcal{F} = \frac{2x_\mathcal{F}}{l_x} - 1, \tag{2}$$

$$\hat{y}_\mathcal{F} = \frac{2y_\mathcal{F}}{l_y} - 1, \tag{3}$$

where the width l_x and height l_y indicate the number of fixation density map pixels in each axis. A value of $\hat{x}_\mathcal{F} = -1$ represents the left-most point and $\hat{x}_\mathcal{F} = 1$ the right-most one. The vertical axis, $\hat{y}_\mathcal{F} = -1$ represents the top-most point and $\hat{y}_\mathcal{F} = 1$ the bottom-most one.

The robot is actuated to look towards the fixation peak. For simplicity, eye movement is assumed to be independent of the exact camera location relative to the playback monitor. For all experiments, only the iCub eyes were actuated while disregarding microsaccadic movements and vergence effects. The positions the iCub should look at are expressed in Cartesian coordinates while assuming the monitor to be at a distance of 30 cm ($\delta = 0.3$) from the image plane. To limit the viewing range of the eyes, $\hat{x}_\mathcal{F}$ and $\hat{y}_\mathcal{F}$ are scaled down by a factor of $\alpha = 0.3$. The Cartesian coordinates are then converted to spherical coordinates by

$$\theta = \arctan\left(\frac{\alpha \cdot \hat{x}_\mathcal{F}}{\sqrt{\delta^2 + (\alpha \cdot \hat{y}_\mathcal{F})^2}}\right), \tag{4}$$

$$\phi = \arctan(\hat{y}_\mathcal{F}), \tag{5}$$

where θ and ϕ are the yaw and pitch angles respectively. These angles are used to actuate the eyes of the iCub such that they pan ∼27° and tilt ∼24° at most[5].

4.3 Experimental Setup

We train the binaural model on a stereo audio-visual dataset and propagate its predicted maps to GASP. We describe the physical setup of the robot environment under which the model used for integrating social cues with binaural sound is evaluated. The human and robot experimental setups closely resemble each other, allowing us to emulate the environmental surrounding experienced by the participants that was described in Sect. 3.2.1.

4.3.1 Binaural Model Training and Evaluation

The binaural DAVE is fine-tuned on a subset of the FAIR-Play dataset [25], comprising 500 randomly chosen videos. The FAIR-Play dataset consists of 1,871 video clips of single or multiple individuals playing musical instruments indoors. Auditory input is binaural with the sound source location maps provided by Wu et al. [75].

Similar to its monaural counterpart, the loss of the binaural DAVE model is computed as Kullback–Leibler divergence between the predicted and ground-truth fixation maps at the last timestep of the 16-frame sequence. The input frames, sound channels and ground-truth maps were together flipped at random during training as an augmentation transform. We use the Adam optimiser with $\beta_1 = .9$, $\beta_2 = .999$, and a learning rate of .001. The model is trained for five epochs with mini-batches containing four sequences of 16 visual frames with their corresponding one-second stereo recordings of audio. We train the model on an NVIDIA GeForce RTX 3080 Ti with 32 GB RAM.

We test our model on 200 randomly chosen clips from the FAIR-Play dataset. Another set of 200 clips are used for validation. Given the close resemblance of audio-visual sound localisation to saliency modelling, we rely on metrics com-

[5] The iCub can pan its eyes within a [−45°, 45°] range and tilt them within a [−40°, 40°] range.

monly used to evaluate the latter [12]. We measure Pearson's correlation coefficient (CC) and similarity (SIM) to quantify the performance of our model. CC calculates the linear correlation between two normalised variables, whereas SIM signifies the similarity between two distributions with a value of 1 indicating that they are identical.

4.3.2 Physical Robot Environment

Some technical adjustments proved necessary to replicate the human experiments on the iCub head as closely as possible. First, the iCub head was placed at a distance of approximately 30 cm from a 24-inch monitor (1920 × 1200 pixel resolution), as depicted in Fig. 1b. This distance is, however, shorter than the 55 cm distance the participants sat from the desktop screen. The distance reduction was performed so that the iCub's field of vision covers a larger portion of the monitor. Since the robot lacks foveated vision, the attention is distributed uniformly to all visible regions, causing the robot to attend to irrelevant environmental changes or visual distractors. Second, the previous robot's eye fixation position needed to be retained as a starting point for the next trial to provide scenery variations to the model. Direct light sources also needed to be switched off to avoid glare. Once the experimental setup was ready, the pipeline started the video playback in fullscreen mode, simultaneously capturing a 30-frame segment of the video using a single iCub camera[6] along with one-second audio recordings from each microphone[7] mounted on the iCub's ears.

In the current study, the iCub head shifts its eyes towards the auditory target. This differs from how participants responded to the stimuli. The participants provided feedback by pressing a key, with their hands were already resting on the keyboard, leading to a much faster response than the time it takes for an iCub head to shift its eyes. This difference could lead to systematic differences in RT, making the RT of the iCub head incomparable to those of the participants. For that reason, the RT of the robot was not measured nor analysed. Nevertheless, it is worth noticing that even though humans and the robot respond differently to a trial, the task they perform is essentially the same. Therefore, ER can be adequately measured and analysed as the robot response. One-way repeated measures ANOVA is used to test the SRC effects of the robot's response under the three congruency conditions (congruent, incongruent and neutral). All post hoc tests in the current study use Bonferroni correction. Additionally, an independent t-test is conducted to compare the difference in SRC effects between humans and the robot. The SRC effect is measured by subtracting congruent responses from incongruent responses.

Table 1 Evaluating the binaural audio-visual sound source localisation model on the test subset of the FAIR-Play dataset

Methods	CC↑	SIM↑
Visual-only DAVE	0.5030	0.3972
Audio-visual DAVE	0.6068	0.4398
Binaural audio-visual DAVE (ours)	**0.6411**	**0.5050**

4.4 Experimental Results

Our binaural audio-visual sound localisation model outperforms monaural and visual-only variants in terms of the CC and SIM metrics. For processing conflicting auditory and visual stimuli, using a binaural model becomes necessary to estimate the direction of sound arrival. This allows us to replicate human-like patterns in attending to sound under congruent, incongruent, and neutral conditions.

4.4.1 Binaural Sound Localisation

We fine-tune the DAVE variants on the FAIR-Play training subset and evaluate the CC and SIM metrics on the test subset. We compare the predicted saliency maps against the ground-truth audio maps for all video frames. The input consists of the preceding 15 frames of a given video's final frame at timestep t_{15} including the final frame. The evaluation results are reported following the fifth training epoch, given that the validation loss increases after that. The binaural DAVE outperforms both the audio-visual and visual-only variants of DAVE, as shown in Table 1.

We observe a significant gap in SIM, but not in CC, between the binaural DAVE and other variants. The SIM metric is highly sensitive to false negatives [12]. Given the objective of localising sounds in the visual stream, saliency prediction models would produce maps uncorrelated with regions having high sound activity. In the case of audio-visual and video-only variants, the models are unaware of the sound location and rely on the activity observed in the visual stream. This implies that those model variants behave like saliency predictors.

In Fig. 5, we observe that the predictions highly correspond to the ground-truth maps, with an incorrect prediction displayed in the last column. Wrong predictions lead to faulty movement on the iCub during inference. We note that such false predictions often occur due to the labels being provided as constant audio maps for entire video clips [75]. Changes during the video in which one musician begins playing at a later stage are ignored, as seen from the example shown in the last column of Fig. 5. As indicated by the hand movement in transition between the timesteps t_0 and t_{15}, the musician is playing the cello.

[6] http://wiki.icub.org/wiki/Cameras

[7] http://wiki.icub.org/wiki/Microphones

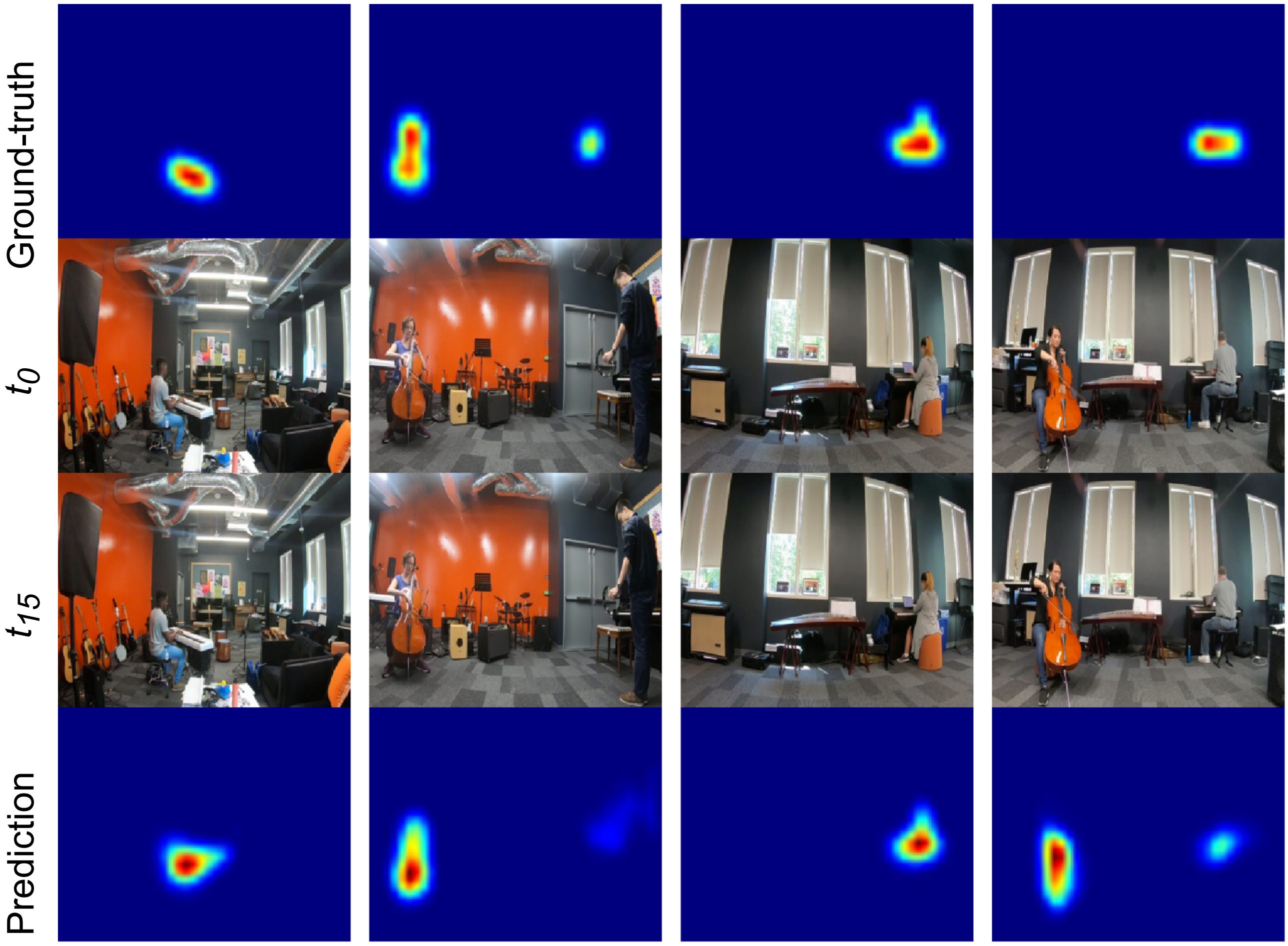

Fig. 5 Qualitative examples showing the binaural DAVE predictions on the FAIR-Play test subset

4.4.2 Error Rates

A repeated measures ANOVA with a Greenhouse-Geisser correction showed that the robot's ER differed significantly between different congruency conditions, $F\ (2, 34) = 8.02$, $p < .01$, $\eta_p^2 = .18$ (see Figs. 2e and f). Post hoc tests showed that the robot presented significantly lower ER under the congruent condition (mean ± SE $= .37 \pm .01$) than the incongruent condition (mean ± SE $= .41 \pm .01$), $p < .01$. However, there was no statistical significance in the difference between the neutral condition (mean ± SE $= .38 \pm .01$) and both other congruency conditions, $p > .05$ in both cases.

4.4.3 Human-Robot Comparison

The SRC effect was computed as the difference between ER under incongruent and congruent conditions ($SRC = ER_{incongruent} - ER_{congruent}$). Results of the t-test displayed that the robot showed a significantly larger SRC effect (mean ± SE $= .04 \pm .001$) than humans (mean ± SE $= .01 \pm .01$), $t\ (72) = 2.35$, $p < .05$ (see Fig. 6).

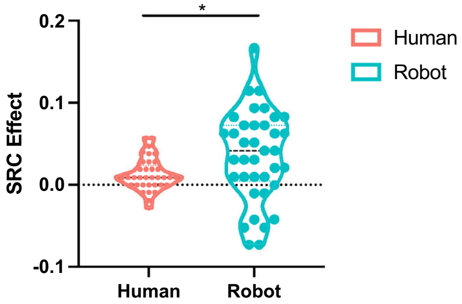

Fig. 6 SRC effects comparison between humans and the robot. $*$ denotes $.01 < p < .05$

5 General Discussion and Conclusions

Our current neurorobotic study investigated human attentional response and modelled the human-like response with the humanoid iCub robot head in a crossmodal audio-visual social attention meeting scenario. According to the research goals, the main findings of the current study are twofold. First, in line with previous crossmodal social attention research [47, 65], our study shows that the visual cue direction enhances the detection of the following auditory target occurring in the same direction, although from a different modality. The current study uses a dynamic gaze shift with corresponding head and upper body movements as visual cue stimuli. It replicates the previous findings by studies using static eye gaze [27, 48], showing a robust reflexive attentional orienting effect. More specifically, the participants show longer RT and higher ER under the incongruent audio-visual condition than the congruent one. Some previous research shows that eye gaze has a stronger attentional orienting effect than simple experimental stimuli (e.g., arrows) [21, 57]. Although we do not have any conditions using arrows as visual cues in our current study, we first demonstrate that realistic and dynamic social cues could have a similar effect in a human crossmodal social attention behavioural study (**H1, H2**). Second, the results from the iCub response demonstrate a successful human-like simulation. With the GASP model, the iCub robot could trigger similar attentional patterns as humans, even in a complex crossmodal scenario. Lastly, the statistical comparison of the SRC effects between humans and the iCub shows that the robot experienced a larger conflict effect than humans (**H3, H4**).

In the human experiment, corresponding to our **H1** hypothesis, social cues that trigger social attention are extended to multiple modalities. Our results support the nature of social gaze cueing, and the view of stimulus-driven shifts of auditory attention might be mediated by other modality information [62]. Furthermore, different from previous gaze-cueing experiments [47], we add a neutral condition to study the interference and facilitation effects during conflict processing. For the neutral condition, participants only see a static meeting scenario without any dynamic social visual cues before the auditory target comes out. The results of humans' RT contradict our **H2** hypothesis. Participants have significantly longer RT under the neutral condition than in the congruent condition. However, no significant difference in RT between neutral and incongruent conditions is found. Thus, the congruent condition in our study has a facilitation effect on the audio-visual conflict processing. These results are consistent with previous studies using the static eye gaze as the visual cue. Their researchers also report a faster response to the gaze-target spatially congruent conditions than the neutral and incongruent conditions, implying a benefit effect of the gaze-oriented attention [20, 60]. The ER results show that the incongruent condition would have significantly more response errors than the congruent condition, and the neutral condition intermediates between incongruent and congruent conditions with slight differences.

In the robot experiment, the iCub experiment results verify **H3** and **H4** that, similarly to humans, the robot's response accuracy is significantly better ($p < .01$) in a congruent condition than in an incongruent one. This similarity is further corroborated by the lack of significant difference ($p > .05$) in both the humans' and the robot's ER in the neutral condition compared to either of the other conditions (cf. Fig. 2c and e). The current study did not directly compare ER between humans and the robot under each condition. Because robots do not respond as accurately as humans, a lower accuracy is to be expected for robots [72]. However, it is still important to find that the relevant values between incongruent and congruent conditions between humans and the robot are closely related. Although the robot shows significantly larger SRC effects than the humans, it is reasonable for responses from the robot to have more variability than those of the humans. Though very low, the iCub's ego noise still makes audio localisation more challenging than for a human who could adjust to the visuals of the avatars in the pretrials. In contrast, the iCub could rely solely on its pretrained model. Besides, although the participants respond to the stimuli by pressing the corresponding keys on the keyboard, while the iCub robot responds by shifting its eyes, the SRC effects still significantly show during the iCub experiment. The robot provides a fixation density map, representing the most likely region a human would tend to fix his/her attention in a crossmodal audio-visual scenario. By providing different degrees of attention to each modality, guaranteeing that all of them would be considered for the determination of the fixation density map, the neurorobotic model is capable of generating the human-like crossmodal attention. The possibility of making a humanoid robot mimic human attention behaviour is an essential step towards creating robots that can understand human intentions, predict their behaviours, and behave naturally and smoothly in human-robot interactions.

6 Future Work

The current work could give way to studies from multiple areas and perspectives. For instance, during the social attention task, eye-tracking techniques could be used to collect human eye movement responses, e.g., pupil dilation, visual fixation, and microsaccades. This allows for a more comprehensive analysis of human attention under the different conditions of audio-visual congruency. Fine-tuning audio-visual saliency models on the collected task-specific data could lead to performance on par with humans.

To make the experimental design more diverse and realistic, future studies could utilise other social cues from the avatar's face and body. Besides, the experimental design could be enhanced by considering additional factors, such as the avatars' emotions and other identity features. This could be helpful for target speaker detection, emotion recognition, and sound localisation in future robotic studies. Considering that speaking activity is one key feature in determining which people to look at [76], it is crucial to consider when creating robots that mimic human attention behaviour. Also, the high performance of the most recent in-the-wild active speaker detection models [13, 36, 58] indicates their reliability in providing accurate attention maps.

Our current work and findings can be applied to build social robots to play with children who have ASD or autistic traits. Previous research has shown that children with ASD avoid mutual gaze and other social interaction with humans, but not with humanoid robots [59]. This can be explained by the fact that humanoid robots with child-like appearance are more approachable by children with ASD [56, 63]. Thus, it is possible and meaningful for social robots to help children with ASD improve their social functions.

Finally, the current experiment could be extended to a human-robot interaction scenario, such as replacing avatars with real humans or robots and evaluating responses from the participants and robots [5]. There have been several human-robot interaction studies about how humans react to a robot's eye gaze [2, 51, 74] or the mutual gaze effect on human decision-making [9, 35]. Based on our study, what can be extended, but can also be challenging, is to make robots learn multiperson eye gaze and detect the active speaker in real-time during a collaborative task or social scenario with humans.

In conclusion, our interdisciplinary study provides new insights into how social cues trigger social attention in a complex multisensory scenario with realistic and dynamic social cues and stimuli. We also demonstrated that by predicting the fixation density map, the GASP model triggered the iCub robot to have a human-like response and similar socio-cognitive functions, resolving sensory conflicts within a high-level social context. By combining stimulus-driven information with internal targets and expectations, we hypothesise that these aspects of multisensory interaction should enable current computational models of robot perception to yield robust and flexible social behaviour during human-robot interaction.

Supplementary Materials

The example of experimental stimuli and videos for both human data and the iCub robot data collection can be viewed at this link: https://www.youtube.com/watch?v=bjiYEs1x-7E.

Acknowledgements We especially thank Dr. Cornelius Weber, Dr. Phillip Allgeuer, Dr. Zhong Yang, Dr. Guochun Yang, and Guangteng Meng for their improvement of the experimental design and the manuscript.

Author Contributions DF, XL, and SW designed the experiment. DF and ZC collected the human data. FA conducted the computational modelling and robotic experiment. DF analysed the data. MK developed the framework for generating the experimental stimuli. MK and ES contributed to the experimental setup and stimuli generation. DF, HC, FA, and MK wrote the manuscript. All authors contributed to improve the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the National Natural Science Foundation of China (NSFC, No. 62061136001), the German Research Foundation (DFG) under project Transregio Crossmodal Learning (TRR 169). D.F. is funded by the Office of China Postdoctoral Council.

Data Availibility The datasets generated and analysed during the current study are available in the Open Science Framework repository, https://osf.io/fbncu/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Informed Consent Informed consent was obtained from all participants in the study.

Ethical Approval All procedures performed in studies involving participants were following the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

1. Abawi F, Weber T, Wermter S (2021) GASP: gated attention for saliency prediction. In: Proceedings of the international joint conference on artificial intelligence (IJCAI), pp. 584–591. IJCAI Organization. https://doi.org/10.24963/ijcai.2021/81
2. Admoni H, Scassellati B (2017) Social eye gaze in human-robot interaction: a review. J Human-Robot Interact 6(1):25–63. https://doi.org/10.5898/JHRI.6.1.Admoni
3. Akiyama T, Kato M, Muramatsu T, Umeda S, Saito F, Kashima H (2007) Unilateral amygdala lesions hamper attentional orienting triggered by gaze direction. Cereb Cortex 17(11):2593–2600. https://doi.org/10.1093/cercor/bhl166
4. Ambrosecchia M, Marino BF, Gawryszewski LG, Riggio L (2015) Spatial stimulus-response compatibility and affordance effects are not ruled by the same mechanisms. Front Hum Neurosci 9:283. https://doi.org/10.3389/fnhum.2015.00283
5. Andriella A, Siqueira H, Fu D, Magg S, Barros P, Wermter S, Torras C, Alenya G (2020) Do I have a personality? Endowing care robots with context-dependent personality traits. Int J Soc Robot. https://doi.org/10.1007/s12369-020-00690-5

6. Montes-y AJST, FA GMG (2019) Gated multimodal networks. Neural Comput Appl 32(14):10209. https://doi.org/10.1007/s00521-019-04559-1
7. Baron-Cohen S (1997) Mindblindness: an essay on autism and theory of mind. MIT press, Cambridge
8. Battich L, Fairhurst M, Deroy O (2020) Coordinating attention requires coordinated senses. Psychonom Bull Rev. https://doi.org/10.3758/s13423-020-01766-z
9. Belkaid M, Kompatsiari K, De Tommaso D, Zablith I, Wykowska A (2021) Mutual gaze with a robot affects human neural activity and delays decision-making processes. Sci Robot 6(58):eabc5044. https://doi.org/10.1126/scirobotics.abc5044
10. Birmingham E, Kingstone A (2009) Human social attention: a new look at past, present, and future investigations. Ann N Y Acad Sci 1156(1):118–140. https://doi.org/10.1111/j.1749-6632.2009.04468.x
11. Brooks R, Meltzoff AN (2005) The development of gaze following and its relation to language. Dev Sci 8(6):535–543. https://doi.org/10.1111/j.1467-7687.2005.00445.x
12. Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F (2019) What do different evaluation metrics tell us about saliency models? IEEE Trans Pattern Anal Mach Intell 41(3):740–757. https://doi.org/10.1109/TPAMI.2018.2815601
13. Carneiro H, Weber C, Wermter S FaVoA: Face-Voice association favours ambiguous speaker detection. In: Proceedings of the 30th international conference on artificial neural networks (ICANN 2021), vol. LNCS 12891:439–450. https://doi.org/10.1007/978-3-030-86362-3_36
14. Cohen JD, Dunbar K, McClelland JL (1990) On the control of automatic processes: a parallel distributed processing account of the stroop effect. Psychol Rev 97(3):332. https://doi.org/10.1037/0033-295x.97.3.332
15. Cornia M, Baraldi L, Serra G, Cucchiara R (2018) Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Trans Image Process 27(10):5142–5154. https://doi.org/10.1109/TIP.2018.2851672
16. Dalmaso M, Zhang X, Galfano G, Castelli L (2021) Face masks do not alter gaze cueing of attention: evidence from the Covid-19 pandemic. I-Perception 12(6):20416695211058480. https://doi.org/10.1177/20416695211058480
17. Doruk D, Chanes L, Malavera A, Merabet LB, Valero-Cabré A, Fregni F (2018) Cross-modal cueing effects of visuospatial attention on conscious somatosensory perception. Heliyon 4(4):e00595. https://doi.org/10.1016/j.heliyon.2018.e00595
18. Eriksen BA, Eriksen CW (1974) Effects of noise letters upon the identification of a target letter in a nonsearch task. Percept Psychophys 16(1):143–149. https://doi.org/10.3758/BF03203267
19. Farroni T, Massaccesi S, Pividori D, Johnson MH (2004) Gaze following in newborns. Infancy 5(1):39–60. https://doi.org/10.1207/s15327078in0501_2
20. Friesen CK, Kingstone A (1998) The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychonom Bull Rev 5(3):490–495. https://doi.org/10.3758/BF03208827
21. Friesen CK, Ristic J, Kingstone A (2004) Attentional effects of counterpredictive gaze and arrow cues. J Exp Psychol Hum Percept Perform 30(2):319. https://doi.org/10.1037/0096-1523.30.2.319
22. Frischen A, Bayliss AP, Tipper SP (2007) Gaze cueing of attention: visual attention, social cognition, and individual differences. Psychol Bull 133(4):694. https://doi.org/10.1037/0033-2909.133.4.694
23. Fu D, Barros P, Parisi GI, Wu H, Magg S, Liu X, Wermter S (2018) Assessing the contribution of semantic congruency to multisensory integration and conflict resolution. In: IROS 2018 Workshop on crossmodal learning for intelligent robotics. IEEE. https://arxiv.org/abs/1810.06748
24. Fu D, Weber C, Yang G, Kerzel M, Nan W, Barros P, Wu H, Liu X, Wermter S (2020) What can computational models learn from human selective attention? A review from an audiovisual unimodal and crossmodal perspective. Front Integr Neurosci 14:10. https://doi.org/10.3389/fnint.2020.00010
25. Gao R, Grauman K (2019) 2.5D visual sound. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), IEEE. pp. 324–333. https://doi.org/10.1109/CVPR.2019.00041
26. Gori M, Schiatti L, Amadeo MB (2021) Masking emotions: face masks impair how we read emotions. Front Psychol 12:1541. https://doi.org/10.3389/fpsyg.2021.669432
27. Guo J, Luo X, Wang E, Li B, Chang Q, Sun L, Song Y (2019) Abnormal alpha modulation in response to human eye gaze predicts inattention severity in children with ADHD. Dev Cogn Neurosci 38:100671. https://doi.org/10.1016/j.dcn.2019.100671
28. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and Imagenet? In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), IEEE. pp. 6546–6555. https://doi.org/10.1109/CVPR.2018.00685
29. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE international conference on computer vision (ICCV), IEEE, USA. pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123
30. He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks. In: European conference on computer vision. Springer. pp. 630–645. https://doi.org/10.1007/978-3-319-46493-0_38
31. Jain S, Yarlagadda P, Jyoti S, Karthik S, Subramanian R, Gandhi V (2020) ViNet: Pushing the limits of visual modality for audio-visual saliency prediction. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE. pp. 3520–3527. https://doi.org/10.1109/IROS51168.2021.9635989
32. Jessen S, Grossmann T (2014) Unconscious discrimination of social cues from eye whites in infants. Proc Natl Acad Sci 111(45):16208–16213. https://doi.org/10.1073/pnas.1411333111
33. Johnson S, Slaughter V, Carey S (1998) Whose gaze will infants follow? the elicitation of gaze-following in 12-month-olds. Dev Sci 1(2):233–238. https://doi.org/10.1111/1467-7687.00036
34. Kerzel M, Wermter S (2020) Towards a data generation framework for affective shared perception and social cue learning using virtual avatars. In: Workshop on affective shared perception, ICDL 2020, IEEE international conference on development and learning https://www.whisperproject.eu/images/WASP2020submissions/9_ICDL_Workshop_WASPKerzelWermter.pdf
35. Kompatsiari K, Ciardo F, Tikhanoff V, Metta G, Wykowska A (2021) It's in the eyes: the engaging role of eye contact in HRI. Int J Soc Robot 13(3):525–535. https://doi.org/10.1007/s12369-019-00565-4
36. Köpüklü O, Taseska M, Rigoll G (2021) How to design a three-stage architecture for audio-visual active speaker detection in the wild. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), IEEE. pp. 1193–1203. https://doi.org/10.1109/ICCV48922.2021.00123
37. Kornblum S, Lee JW (1995) Stimulus-response compatibility with relevant and irrelevant stimulus dimensions that do and do not overlap with the response. J Exp Psychol Hum Percept Perform 21(4):855. https://doi.org/10.1037//0096-1523.21.4.855
38. Langton SR, Watt RJ, Bruce V (2000) Do the eyes have it? Cues to the direction of social attention. Trends Cogn Sci 4(2):50–59. https://doi.org/10.1016/s1364-6613(99)01436-9
39. Laube I, Kamphuis S, Dicke PW, Thier P (2011) Cortical processing of head-and eye-gaze cues guiding joint social attention. Neu-

roimage 54(2):1643–1653. https://doi.org/10.1016/j.neuroimage.2010.08.074

40. Liu X, Liu T, Shangguan F, Sørensen TA, Liu Q, Shi J (2018) Neurodevelopment of conflict adaptation: evidence from event-related potentials. Dev Psychol 54(7):1347. https://doi.org/10.1037/dev0000524
41. MacLeod CM (1991) Half a century of research on the stroop effect: an integrative review. Psychol Bull 109(2):163. https://doi.org/10.1037/0033-2909.109.2.163
42. Maddox RK, Pospisil DA, Stecker GC, Lee AK (2014) Directing eye gaze enhances auditory spatial cue discrimination. Curr Biol 24(7):748–752. https://doi.org/10.1016/j.cub.2014.02.021
43. McNeely HE, West R, Christensen BK, Alain C (2003) Neurophysiological evidence for disturbances of conflict processing in patients with schizophrenia. J Abnorm Psychol 112(4):679. https://doi.org/10.1037/0021-843X.112.4.679
44. Mundy P, Newell L (2007) Attention, joint attention, and social cognition. Curr Dir Psychol Sci 16(5):269–274. https://doi.org/10.1111/j.1467-8721.2007.00518.x
45. Newport R, Howarth S (2009) Social gaze cueing to auditory locations. Q J Experiment Psychol 62(4):625–634. https://doi.org/10.1080/17470210802486027
46. Nocentini O, Fiorini L, Acerbi G, Sorrentino A, Mancioppi G, Cavallo F (2019) A survey of behavioral models for social robots. Robotics 8(3):54. https://doi.org/10.3390/robotics8030054
47. Nuku P, Bekkering H (2008) Joint attention: inferring what others perceive (and don't perceive). Conscious Cogn 17(1):339–349. https://doi.org/10.1016/j.concog.2007.06.014
48. Nuku P, Bekkering H (2010) When one sees what the other hears: crossmodal attentional modulation for gazed and non-gazed upon auditory targets. Conscious Cogn 19(1):135–143. https://doi.org/10.1016/j.concog.2009.07.012
49. Nummenmaa L, Calder AJ (2009) Neural mechanisms of social attention. Trends Cogn Sci 13(3):135–143. https://doi.org/10.1016/j.tics.2008.12.006
50. Parisi GI, Barros P, Fu D, Magg S, Wu H, Liu X, Wermter S (2018) A neurorobotic experiment for crossmodal conflict resolution in complex environments. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE. pp. 2330–2335. https://doi.org/10.1109/IROS.2018.8594036
51. Pfeifer-Lessmann N, Pfeifer T, Wachsmuth I (2012) An operational model of joint attention-timing of gaze patterns in interactions between humans and a virtual human. In: Proceedings of the annual meeting of the cognitive science society, vol. 34. https://escholarship.org/uc/item/4f49f71h
52. Posner M, Cohen Y (1984) Components of visual orienting. Attention and performance X: Control of language processes. Psychology Press, London, pp 531–556
53. Posner MI, Snyder CR, Davidson BJ (1980) Attention and the detection of signals. J Exp Psychol Gen 109(2):160. https://doi.org/10.1037/0096-3445.109.2.160
54. Proctor RW, Vu KPL (2006) Stimulus-response compatibility principles: data, theory, and application. CRC Press, Cambridge
55. Rachavarapu KK, Sundaresha V, Aakanksha Rajagopalan A (2021) Localize to binauralize: Audio spatialization from visual sound source localization. In: Proceedings of the IEEE/cvf international conference on computer vision, IEEE. pp. 1930–1939. https://doi.org/10.1109/ICCV48922.2021.00194
56. Raptopoulou A, Komnidis A, Bamidis PD, Astaras A (2021) Human-robot interaction for social skill development in children with Asd: a literature review. Healthcare Technol Lett 8(4):90–96. https://doi.org/10.1049/htl2.12013
57. Ristic J, Wright A, Kingstone A (2007) Attentional control and reflexive orienting to gaze and arrow cues. Psychonom Bull Rev 14(5):964–969. https://doi.org/10.3758/bf03194129
58. Roth J, Chaudhuri S, Klejch O, Marvin R, Gallagher A, Kaver L, Ramaswamy S, Stopczynski A, Schmid C, Xi Z, Pantofaru C (2020) AVA-ActiveSpeaker: An audio-visual dataset for active speaker detection. In: Proceedings of the IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE. pp. 4492–4496. https://doi.org/10.1109/ICASSP40776.2020.9053900
59. Scassellati B, Admoni H, Matarić M (2012) Robots for use in autism research. Ann Rev Biomed Eng 14:275–294. https://doi.org/10.1146/annurev-bioeng-071811-150036
60. Schuller AM, Rossion B (2004) Perception of static eye gaze direction facilitates subsequent early visual processing. Clin Neurophysiol 115(5):1161–1168. https://doi.org/10.1016/j.clinph.2003.12.022
61. Senju A, Johnson MH (2009) Atypical eye contact in autism: models, mechanisms and development. Neurosci Biobehav Rev 33(8):1204–1214. https://doi.org/10.1016/j.neubiorev.2009.06.001
62. Shepherd SV (2010) Following gaze: gaze-following behavior as a window into social cognition. Front Integr Neurosci 4:5. https://doi.org/10.3389/fnint.2010.00005
63. Shimaya J, Yoshikawa Y, Matsumoto Y, Kumazaki H, Ishiguro H, Mimura M, Miyao M (2016) Advantages of indirect conversation via a desktop humanoid robot: Case study on daily life guidance for adolescents with autism spectrum disorders. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN), IEEE. pp. 831–836. https://doi.org/10.1109/ROMAN.2016.7745215
64. Simon JR, Rudell AP (1967) Auditory SR compatibility: the effect of an irrelevant cue on information processing. J Appl Psychol 51(3):300. https://doi.org/10.1037/h0020586
65. Soto-Faraco S, Sinnett S, Alsius A, Kingstone A (2005) Spatial orienting of tactile attention induced by social cues. Psychonom Bull Rev 12(6):1024–1031. https://doi.org/10.3758/BF03206438
66. Sperdin HF, Coito A, Kojovic N, Rihs TA, Jan RK, Franchini M, Plomp G, Vulliemoz S, Eliez S, Michel CM, Schaer M (2018) Early alterations of social brain networks in young children with autism. ELife 7:1–23. https://doi.org/10.7554/eLife.31670
67. Srinivasan SM, Eigsti IM, Neelly L, Bhat AN (2016) The effects of embodied rhythm and robotic interventions on the spontaneous and responsive social attention patterns of children with autism spectrum disorder (Asd): a pilot randomized controlled trial. Res Autism Spect Disord 27:54–72. https://doi.org/10.1016/j.rasd.2016.01.004
68. Stajduhar A, Ganel T, Avidan G, Rosenbaum RS, Freud E (2022) Face masks disrupt holistic processing and face perception in school-age children. Cogn Res Princ Implic 7(1):1–10. https://doi.org/10.1186/s41235-022-00360-2
69. Stroop JR (1935) Studies of interference in serial verbal reactions. J Exp Psychol 18(6):643.https://doi.org/10.1037/h0054651
70. Tavakoli HR, Borji A, Kannala J, Rahtu E (2020) Deep audio-visual saliency: Baseline model and data. In: ACM symposium on eye tracking research and applications, ETRA '20 Short Papers. Association for Computing Machinery, New York, NY, USA. pp. 1–5. https://doi.org/10.1145/3379156.3391337
71. Tsiami A, Koutras P, Maragos P (2020) STAViS: Spatio-temporal audiovisual saliency network. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition (CVPR), IEEE. pp. 4766–4776. https://doi.org/10.1109/CVPR42600.2020.00482
72. Wang J, Wang J, Qian K, Xie X, Kuang J (2020) Binaural sound localization based on deep neural network and affinity propagation clustering in mismatched HRTF condition. EURASIP J Audio Speech Music Process 2020(4):1–16. https://doi.org/10.1186/s13636-020-0171-y

73. Wightman FL, Kistler DJ (1997) Monaural sound localization revisited. J Acoust Soc Am 101(2):1050–1063. https://doi.org/10.1121/1.418029
74. Willemse C, Marchesi S, Wykowska A (2018) Robot faces that follow gaze facilitate attentional engagement and increase their likeability. Front Psychol 9:70. https://doi.org/10.3389/fpsyg.2018.00070
75. Wu X, Wu Z, Ju L, Wang S (2021) Binaural Audio-Visual Localization, vol. 35(4). AAAI. https://doi.org/10.1609/aaai.v35i4.16403
76. Xu M, Liu Y, Hu R, He F (2018) Find who to look at: turning from action to saliency. IEEE Trans Image Process 27(9):4529–4544. https://doi.org/10.1109/TIP.2018.2837106
77. Yeung HH, Werker JF (2013) Lip movements affect infants' audiovisual speech perception. Psychol Sci 24(5):603–612. https://doi.org/10.1177/0956797612458802

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.