A developmental model of audio-visual attention (MAVA) for bimodal language learning in infants and robots

Raphaël Bergoin, Sofiane Boucenna, Raphaël D'urso, David Cohen, Alexandre Pitti

scientific reports

Check for updates

OPEN

A developmental model of audio-visual attention (MAVA) for bimodal language learning in infants and robots

Raphaël Bergoin[1], Sofiane Boucenna[1]✉, Raphaël D'Urso[1], David Cohen[2,3] & Alexandre Pitti[1]

A social individual needs to effectively manage the amount of complex information in his or her environment relative to his or her own purpose to obtain relevant information. This paper presents a neural architecture aiming to reproduce attention mechanisms (alerting/orienting/selecting) that are efficient in humans during audiovisual tasks in robots. We evaluated the system based on its ability to identify relevant sources of information on faces of subjects emitting vowels. We propose a developmental model of audio-visual attention (MAVA) combining Hebbian learning and a competition between saliency maps based on visual movement and audio energy. MAVA effectively combines bottom-up and top-down information to orient the system toward pertinent areas. The system has several advantages, including online and autonomous learning abilities, low computation time and robustness to environmental noise. MAVA outperforms other artificial models for detecting speech sources under various noise conditions.

Social interactive agents have to manage large amounts of information in complex and changing environments. For example, infants must address noise and uncertainty during their first social interactions with caregivers. Given the difficulty of analyzing large amounts of information simultaneously, agents must determine what information is spatially and temporally relevant. In humans, the brain has limited capacity to process all sensory stimuli present in the physical world at any point in time and relies on the cognitive process of attention to focus neural resources on the most relevant information[1]. According to Posner, attention includes three basic functions: maintaining the alert state; orienting to sensory stimuli; and selecting the target stimuli among competing sources or responses (executive functions)[2]. For example, infants shift their attention from the eyes to the mouth between 4 and 8 months. Initially, infants tend to naturally orient their gaze toward the eyes. Then, at the pivotal age for language development, infants shift more their attention toward the mouth and focus on the movements of the speaker's mouth[3]. This capability is crucial for language development, as it helps children associate sounds with their corresponding visual cues. For example, Kuhl et al. demonstrated that when a newborn was placed in front of a screen with 2 images of an adult's face producing 2 different vowels ("a", "e", "i", "o", or "u") and a speaker was placed behind the newborn emitting one of the 2 vowel sounds, the newborns preferentially turned toward the face images that were congruent with the emitted sound, and the infants tended to mimic the perceived vowels[4]. These observations show that continuous and multimodal learning mechanisms are applied during infant development. Some studies have also shown that infants can match stimuli perceived by different sense modalities; however, the mechanisms underlying these abilities are unclear[5].

From a computational point of view, two main approaches have been developed to model an agent's attention in the visual space. Notably, these methods do not necessarily simulate the development of infant attention centered on the face. First, we distinguish different probabilistic and statistical techniques, such as Bayesian surprise[6–8], canonical correlations[9–11] and mutual information[12,13]. Although these approaches have obtained interesting results and show a certain robustness to noise, these statistical models typically use paradigms with considerable computational costs, and a large amount of data is needed to guarantee consistent statistical results. In contrast, bioinspired approaches typically use competing saliency maps to mimic neural networks, with the different visual features integrated into topographical saliency maps[14–17]. Itti et al. proposed a model inspired by the neural architecture of the early primate visual system. Their system considers the complex problem of scene

[1]ETIS, UMR 8051, ENSEA, CY Cergy Paris Université, CNRS, Cergy-Pontoise, France. [2]Service de Psychiatrie de l'Enfant et de l'Adolescent, Hôpital Pitié-Salpêtrière, AP-HP, Paris, France. [3]Institut des Systèmes Intelligents et de Robotiques, Université Pierre et Marie Curie, Paris, France. ✉email: sofiane.boucenna@cyu.fr

understanding by rapidly and efficiently selecting locations that should be analyzed in detail[16]. This approach can be used to reproduce bottom-up attention mechanisms and is consistent with feature integration theory, which claims that a few simple visual feature dimensions are represented in the early stages of cortical visual processing (e.g., color, motion, edge orientation)[18]. However, these models do not explain facial attention. The selection process can also be formulated with a bottom-up approach, with the selection process driven by other factors, such as emotional stimuli (e.g., happy faces). In contrast to bottom-up selection, top-down selection implies that selection is completely controlled by an observer's will (a person can choose at will what to select in the environment)[1,19–26]. Posner illustrated two-way selection by proposing an experimental approach called the endogenous cueing procedure[27]. Bottom-up selection is determined by the properties of the features present in the environment and occurs in a passive automatic way, whereas top-down selection is an active and intentional process[1,28–31].

Overall, the attention that infants pay to speakers' mouth movements is a critical aspect of language development and provides a foundation for the development of more complex language skills. In the present study, we propose to model this mechanism using robots as tools to investigate cognitive models[32,33]. Our goal is to link developmental science with robotics through interdisciplinary insights[34]. This approach aims to (1) deepen our understanding of higher human cognitive functions using a synthetic methodology; (2) implement learning mechanisms in robots to better understand children's cognitive development; and (3) enable the development of robots capable of adapting more effectively to their environment. Firstly, we propose a developmental model based on audio-visual attention (*MAVA*) to capture relevant information in an audiovisual task, thereby "mimicking" infants' behaviors, including their early ability to focus on the human mouth. *MAVA* is a minimal neural model that combines Hebbian learning with saliency maps from visual movement and audio energy detection. It also uses complementary and temporally synchronized auditory and visual speech signals originating from the speaker's mouth. Secondly, we designed several experiments to test the ability of the proposed model to locate the mouth during the visualization of subjects pronouncing vowels, including in environments with external visual and noise perturbations. Compared to other architectures, *MAVA* demonstrated how robustness can be achieved. For this purpose, the attentional mechanism must operate autonomously (unsupervised), online, and during robot interactions. Finally, we discuss the results of the neural model from a developmental perspective, specifically comparing *MAVA* with Posner's attention model. This comparison highlights how *MAVA*'s mechanisms align with established theories of attention underlying infants' focus on speech-related visual cues.

Methods

Architecture

Our architecture called *MAVA* is based on two main aspects. First, *MAVA* is based on Posner's model of attention, which distinguishes three basic functions: alerting, orienting, and selecting. Second, to model a system based on human attention as faithfully as possible, we designed an architecture based on the competition among different saliency maps. More precisely, in the proposed architecture, multimodal synchrony is used as top-down feedback. When two events are temporally and spatially synchronized, they are likely to be correlated or at least perceived as correlated[35].

Figure 1 presents MAVA's architecture based on audio-visual attention. It allows for mimicking infants' behaviors, such as focusing on the human mouth. *MAVA* is based on three components: (1) sensory inputs which are movement detection and audio energy, (2) hebbian learning enables the association of the two modalities to identify which area of the image is synchronized with the audio signal and (3) selection mechanism allowing to focus on the most active region. *MAVA* was trained to associate movement in an image (i.e., a moving

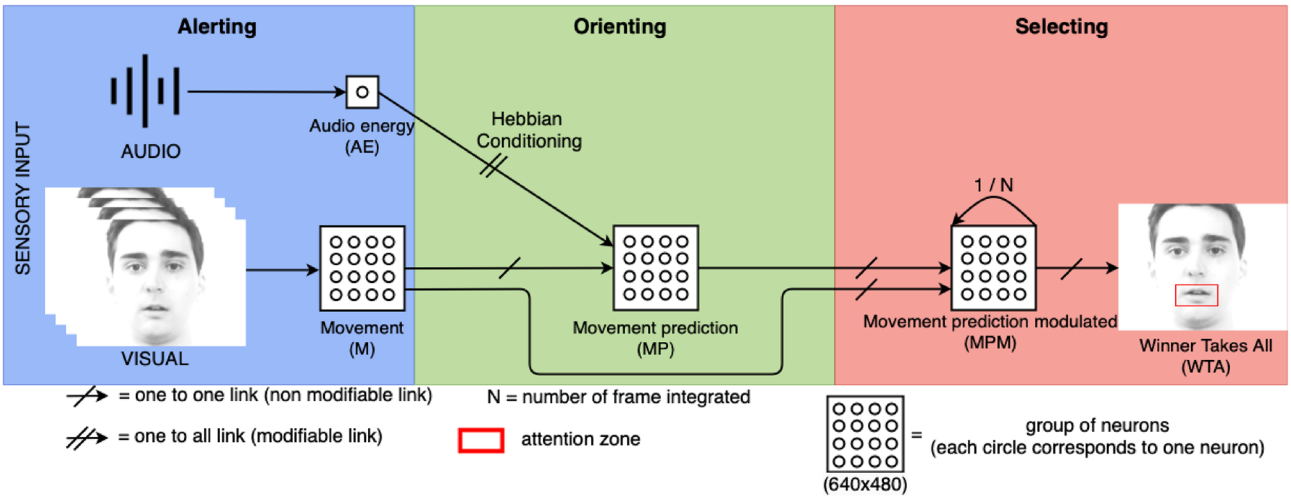

Figure 1. Architecture of the developmental model of audio-visual attention (*MAVA*). The one-to-one link corresponds to a step-by-step multiplication process. The one-to-all link corresponds to a scalar multiplication process. The red rectangle in the right image represents the attention zone centered on the most activated pixel. *The source of human figures belongs to my students..

human face) with the sound produced, allowing the model to detect the correlations between the audio and visual modalities. Consequently, *MAVA* can predict the visual modality as a function of the auditory modality. Movement detection (M) involves obtaining a visual saliency map and considering the bottom-up stimuli to determine the relevant areas in the visual field. This approach involves calculating the optical flow between two consecutive images (only the magnitude of the optical flow is considered because only the intensities at the different locations are useful to *MAVA*)[21,36].

The sound is the stimulus used to calculate the audio energy (*AE*) with the root-mean-square formula, as described in (1), where a_i represents the audio data and N_c indicates the number contained in each data interval.

$$AE = \sqrt{\frac{\sum_{i=1}^{N_c} a_i^2}{N_c}} \tag{1}$$

Then, a Hebbian learning rule is used to associate the *AE* with *M* in the visual scene. Thus, when the two modalities are activated simultaneously, the neural connections are reinforced. The learning rule and the activity of a computational neuron are described in Eqs. (3) and (2):

$$MP_i = w_i(t).AE \tag{2}$$

$$w_i(t+1) = w_i(t) + \epsilon.AE.M_i, \tag{3}$$

where w_i represents the weight of neuron i, ϵ denotes the learning rate, *AE* represents the audio energy, M_i is the movement associated with neuron i and MP_i is the movement prediction based on the output activity of neuron i.

The network output (MP) corresponds to the top-down stimuli that predict the location of the sound in the image and is presented as an activity map. Then, the two outputs (MP and M) compete according to Eq. (4):

$$MPM_i(t+1) = M_i(t) * MP_i(t) + \frac{1}{N} * MPM_i(t) \tag{4}$$

The resulting output, called motion prediction modulation (MPM), focuses on moving areas of the image. MPM is then time-integrated by weighted averaging. This process refines the focus on dynamic image elements, improving the accuracy of sound location prediction. Ultimately, this temporal integration enhances synchronization between visual and auditory signals. Finally, a decision is determined with the winner-takes-all (*WTA*) approach based on the *MPM* to identify the attention zone.

Evaluation protocol

Comparison models

We compared MAVA with two other models. First, we consider movement detection alone, corresponding to bottom-up stimuli. This approach allows us to determine if our neural model shows better performance than movement detection alone and filter the noise in the visual scene. As a second comparison model, we implemented the mutual information technique used by Rolf et al.[13] for similar attention tasks. This model, which is based on the theory of probability, represents the degree of statistical dependence of two random variables. It measures the synchrony between two stimuli in the spatiotemporal domain according to their rhythm and intensity[37]. In our case, this statistical model calculates the mean and variance of the image pixels and audio energy and updates these values over time (for all frames of a data sample). Then, the model expresses the degree of synchrony between the audio and video data in terms of the mutual information. If the pixel variance is less than a specified threshold, the mutual information is set to zero to remove extreme values. Finally, morphological erosion filtering is performed to eliminate excess noise, and a winner-takes-all approach is applied to identify the focus point. The aim is to compare *MAVA* with a statistical approach based on a technique that has been proven efficient for similar tasks[12,37,38]. Moreover, this technique appears to outperform other existing synchrony measures[38].

Noise conditions

To evaluate the robustness of all the techniques implemented, three visual noise conditions (Fig. 2) and one audio noise condition are tested.

Condition 1: In the normal reference condition, no noise is added to the images.
Condition 2: In the condition with salt-and-pepper noise added to images, some pixels in the image are randomly changed to black or white (with a probability of 1%). This kind of noise is often present in image processing and introduces multiple high-intensity distractors, inducing considerable movement in various regions, which may have a strong impact on algorithm performance.
Condition 3: To simulate more realistic visual noise, a second noise condition is considered. We added object noise (a circle of 4 colors) to the upper left corner of each visual frame, with 3 pixels moved diagonally from one frame to another. This introduces a large area of high intensity movement since the noise remains concentrated in the same spot (whereas in the other condition, the noise was purely random and more scattered). A realistic interaction between the robot and the human is simulated, with interaction distances being less than one meter. In this context, the distractor aims to truly attract attention, similar to a person passing through the visual field. Consequently, the noise from the object is placed outside the face.
Condition 4: A percentage of noise is added to the audio signal to simulate the robot's ability to hear the emitted sound or not. This allows testing the algorithm's robustness under extreme auditory conditions.

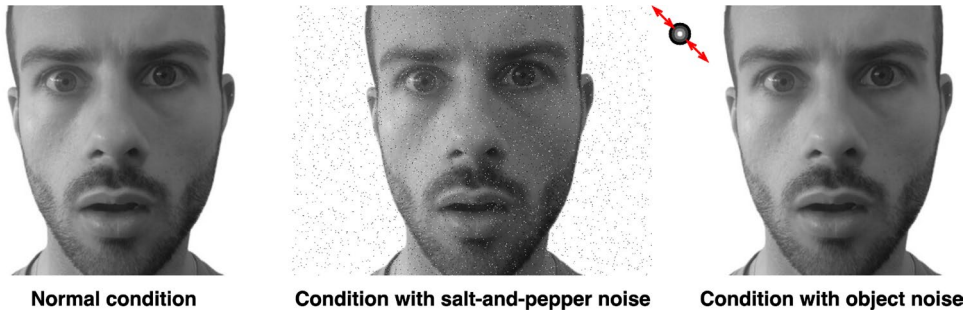

Figure 2. Representation of the three visual test conditions for the same video frame. The first image shows the normal condition without any noise. The second image shows the condition with salt-and-pepper noise. The third image shows the condition with object noise, which is implemented in the upper left corner. *The source of human figures is Raphaël Bergoin (co-author of the paper)..

Type of learning

In our experiment, participants imitate vowels. The complexity of the auditory input has no effect because MAVA only uses audio energy. Thus, whether the participant pronounces phonemes, words, or sentences does not impact the attention mechanism. MAVA only considers the audio energy. To demonstrate the performance of MAVA, we evaluate its ability to identify the sound source in the visual space.

Offline learning: Initially, the experimental protocol is divided into two stages: a training stage and a testing stage. The first stage sets up the Hebbian network for the next step; however, we later note that this stage is not necessary. The goal of the neural network is to learn to associate the sound energy with the movement in an image. To achieve this, the neural network is trained based on 50% of the database (2 samples for each vowel and each subject). Thus, the model updates its weights based on the M and AE values calculated for each frame for each sample. In the testing stage, we use the remaining 50% of the database and determine for each sample if the attention zone is located in the mouth area, as shown in Fig. 3. If the focus point (pixel with the highest activity) is in the rectangle (fixed position), the performance of the attention mechanism is validated for this sample; otherwise, the attention mechanism is determined to be invalid. The same testing and validation processes are performed with the comparison models.

The database is composed of data obtained from 40 different adults, each emitting 5 vowel sounds ('A', 'E', 'I', 'O', and 'U') 4 times, yielding a total of 800 samples. Each sample is composed of 6 video frames of size 640×480 and 6 audio frames of 0.25 seconds.

Online learning: The second experimental protocol consists of combining the training and testing stages by setting up an online learning paradigm. Each frame in the database is input to the system for training, and the results are directly evaluated in the same manner as discussed above. Thus, the experiment can be performed in real time with a robot, with the robot processing the visual and audio information of the subject facing it. Thus, we can assess the adaptability of the robot in focusing its attention on the information area.

Results

Offline learning

The results are presented in Fig. 4. As discussed in the presentation of evaluation protocol, these results are obtained after the training step.

For all vowels and all noise conditions, *MAVA* outperforms the control models, with a focalization rate of 95% (Fig. 4a–c). The different kinds of visual stimuli had a negligible effect on model performance (Fig. 4d). Under the normal condition (i.e., without any noise), we observed that the statistical model obtains mixed

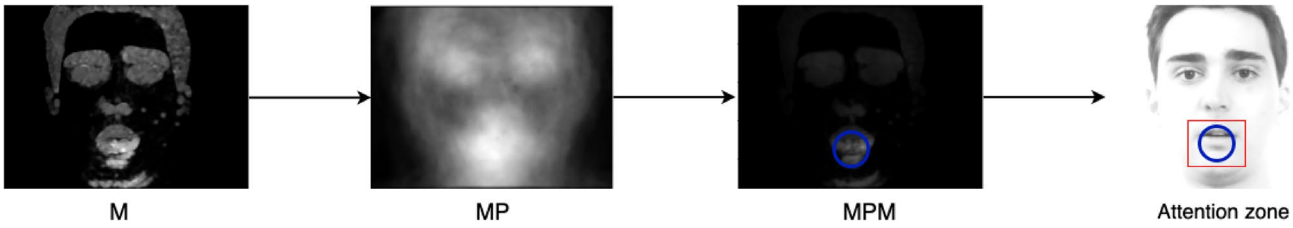

Figure 3. Robot visual attention during an interaction with a human. The different blocks represent each group of neurons (M, MP, and MPM). More white pixels in the image indicate that the corresponding neurons are more active. The blue circle represents the attention zone. The attention zone is identified in the last frame of a sample by determining the highest activated pixel in the MPM. The blue cross symbolizes the point of focus (where the decision is made). The red rectangle represents the theoretical mouth area, which is used to validate the performance of the attention mechanism. *The source of human figures belongs to my students..

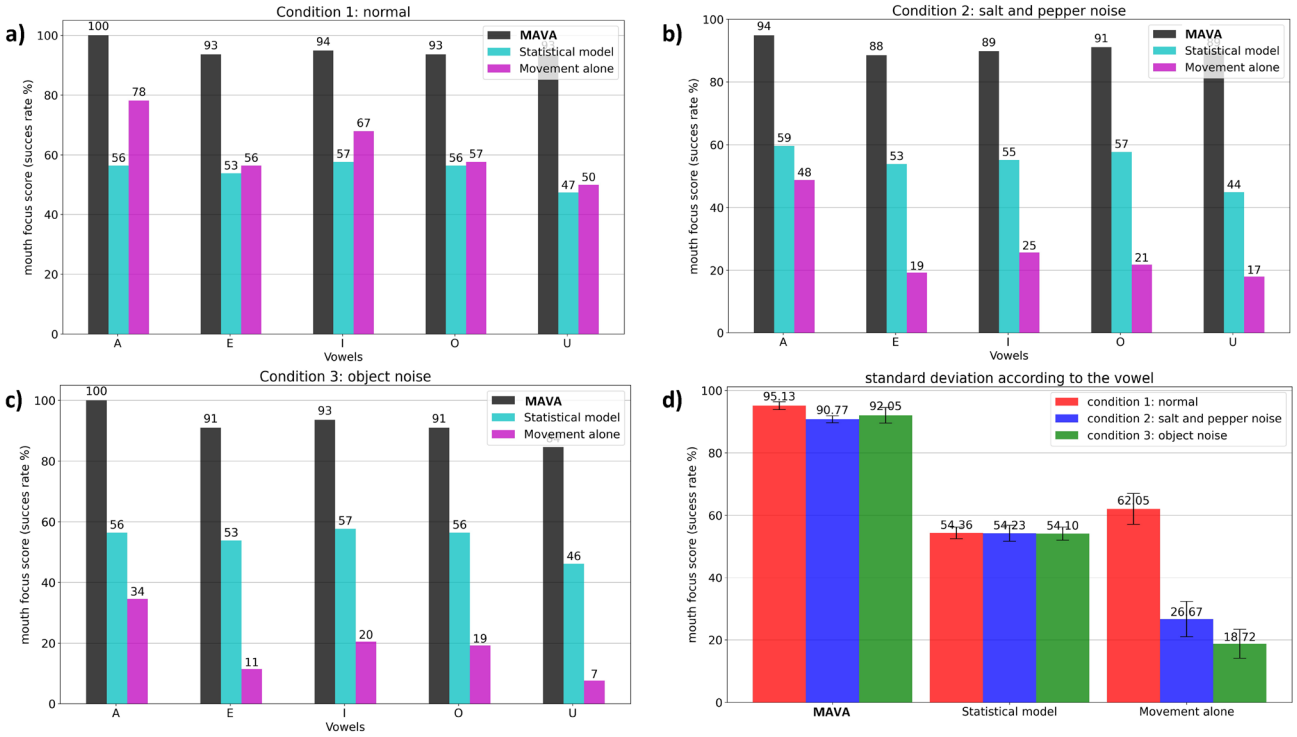

Figure 4. Percent of the attention on the mouth area for each model in visual condition. (**a**) Per vowels in the normal condition, (**b**) per vowels in the salt-and-pepper visual noise condition, (**c**) per vowels in the object visual noise condition and (**d**) average for all noise conditions. The standard deviation (represented by the black brackets) is calculated according to the vowels.

performance (54% focalization on the mouth). Moreover, this model appears to work well for some subjects but performs poorly for other subjects, as demonstrating by its higher standard deviation (Fig. 4d). However, the type of vowel does not seem to have a particular impact on this model (Fig. 4a). In addition, the average performance of the model using only movement detection (61%) is better than that of the statistical model but much inferior to that of *MAVA*.

In summary, *MAVA* handles noise conditions (salt-and-pepper noise condition and object noise condition) and focuses on the speakers' mouth better than the other models (Fig. 4). *MAVA* can compensate for the noise introduced into the images. *MAVA* does not take the shape of object noise into account. The shape of the object has no impact on *MAVA* because it relies only on movement in the image. The results also show the percentage of attention on the mouth area as a function of visual noise (Fig. 5) and auditory noise (Fig. 6). The results indicate that MAVA is more robust to visual and auditory noise than other models, demonstrating MAVA's effectiveness in disturbed environments. The statistical model handles the visual and audio noise well but obtains lower scores than *MAVA*. The movement model is significantly impacted by the salt-and-pepper noise condition because the

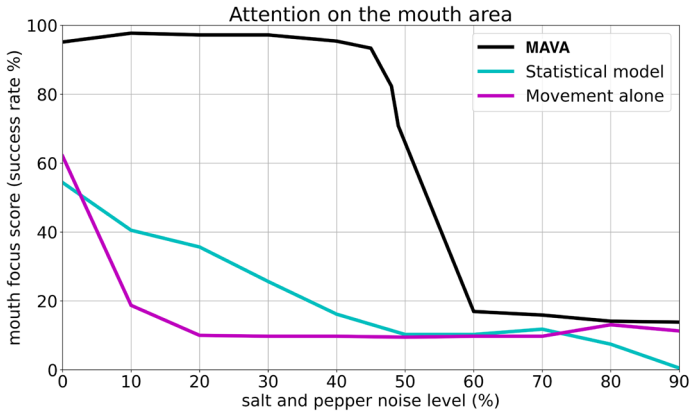

Figure 5. Percent of the attention on the mouth area according to the visual noise level (condition 2).

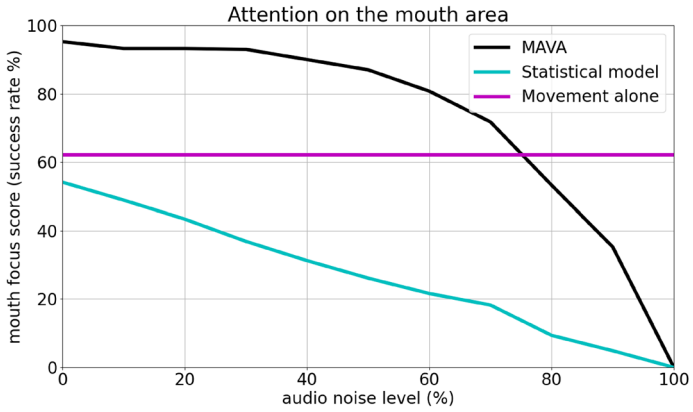

Figure 6. Percent of the attention on the mouth area according to the audio noise level (condition 4).

noise adds strong movement activities throughout the visual space. In the object noise condition, a movement in a specific area attracts more attention than movements in the lip region.

Online learning

For online learning, we tested our model under the same normal and noise conditions. Thus, we can represent the evolution of learning based on the entire database in terms of how well the model focuses on the mouth area (Fig. 7).

MAVA performs similarly under the three conditions, with slightly faster learning observed for the normal condition and condition 3 (object noise). The performance is worse at the beginning of training; however, the system converges quickly (approximately half of the dataset used for training under each condition) to a rate of approximately 95%. After the training phase, the model appears to converge to a similar final rate regardless of the noise condition, and the model approaches the performance obtained during offline learning.

Because the model performs effective online learning, real-time applications can be considered. In the robot experiment, the system can quickly focus its attention on the mouth of the speaker at different places in the visual space. This shows that our architecture can operate in real-world environments and adapt to changes in the visual field.

Discussion and conclusion

The results show that *MAVA* can identify the most relevant area of attention in experiments with talking faces. *MAVA* is based on movement detection, audio energy and Hebbian learning ("cells that fire together, wire together")[39], leading to the rapid convergence of the model. The results show that offline and online learning lead to similar performance since the model converges quickly during training. Consequently, *MAVA* can

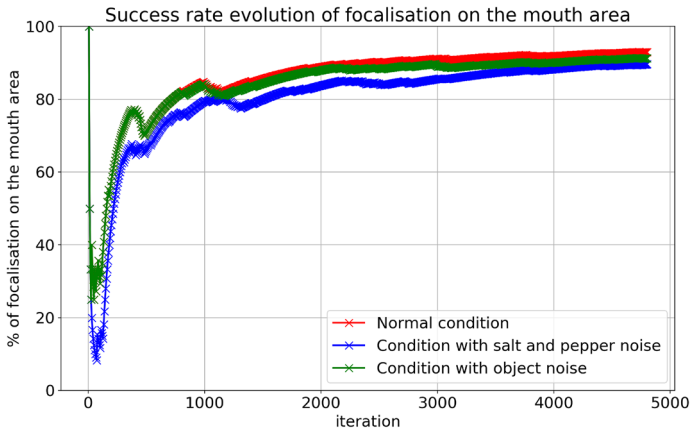

Figure 7. Evolution of the success rate of focalization on the mouth area based on the entire database during the training. Because training was performed online, a measurement (symbolized by the crosses) is made based on each sample (every 6 frames/iterations). In this case, the rate is determined by calculating the moving average of the performance of the tested samples. The red crosses symbolize the measurements obtained under normal conditions, the blue crosses symbolize the measurements obtained under salt-and-pepper noise conditions, and the green crosses symbolize the measurements obtained under object noise conditions.

adapt quickly to disturbances, is robust to different noise and has good properties for robotics applications. In all conditions, the results show that *MAVA* outperforms the other models (*MAVA* is 40% better) and that the participants do not have a significant impact on the results. The statistical model is less robust for two reasons: (1) participants tend to move other parts of their face (eyebrows, eyelids, etc.) when they emit a sound, and (2) although postprocessing leads to noise resistance, the parameters must be properly adjusted. The movement detection model shows poor performance because it depends on the subjects and vowels emitted. The results show that the movement alone (bottom-up selection) model is insufficient for this type of experiment.

From a cognitive science perspective, MAVA follows Posner's model of attention (Fig. 1). Posner distinguishes three basic functions for attention: (1) alerting, which means staying awake and remaining alert to changes in the environment, which corresponds to the sensory input $AE + M$ in *MAVA*; (2) orienting toward significant stimuli, which corresponds to movement prediction based on links with the sensory input through Hebbian learning; and (3) selecting among competing sources or responses, which corresponds to *MPM* saliency derived from movements and movement predictions[2,27,40]. In Posner's model, selection among competing sources occurs mainly through executive functions, specifically inhibition[41]. In *MAVA*, selection is performed based on synchrony. Interestingly, executive functions are not mature at birth and develop slowly during development. Inhibition is one of the last functions to mature during development[41,42]. Therefore, inhibition may not be a unique mechanism, even if it becomes dominant during childhood. Similar to the results of *MAVA*, several studies have noted that synchrony may be a signal that guides attention during language development[43,44] or imitation[45,46], allowing the detection of amodal invariants[43–45]. Another study showed that visual and tactile maps can be aligned through a Hebbian learning stage to produce emergent properties, such as sensitivity toward the spatial configuration of the face, which can be used to detect eye and mouth movement[47].

In its present form, *MAVA* has some limitations, and future work should focus on three aspects. First, the agent/robot should be allowed to switch from one target to another. To achieve this, we need to model an inhibition mechanism. As previously noted, inhibition is needed to process selective attention, and it allows the brain to trigger a cognitive control mechanism[42]. For example, sustained visual attention on the mouth will trigger inhibition in the mouth area, allowing the robot to focus on other areas, such as the eyes. However, the prefrontal cortex, which supports inhibition processes in adults, is immature in infants[48]. This maturity is compatible with the importance of attention in learning and communicating. Inhibition is an important mechanism that influences predictions in people with attention disorders. For example, children with autism spectrum disorder have difficulties inhibiting their attention when a social stimulus is encountered. This may lead to difficulties in social and language development[28]. The second aspect is to use *MAVA* to improve audio-visual vowel recognition. An artificial neural model that simulates Kuhl's experiment should be developed[4]. This experiment showed that infants look at human faces that are congruent with the sound emitted by a loudspeaker. Our architecture can be combined with artificial neural networks to (i) extract visual and audio features, (ii) alter the internal state of the robot's motor and (iii) introduce an attention mechanism that can focus on the mouth. This would show the impact of *MAVA* in the early stages of language development. The last aspect is to test the hypothesis that bilingual infants have more difficulties perceiving emotional cues because they focus more on the mouth[49]. Therefore, prolonged attention to the mouth may affect the development of social communication abilities, particularly facial expressions[50]. Attention to the eye region appears to facilitate the development of facial expertise in infants[51,52]. We will integrate MAVA into our sensory-motor architecture, enabling the model to recognize facial expressions[45] and analyze whether prolonged attention to the mouth affects the learning and recognition of these expressions. Additionally, we plan to employ MAVA in a robotics task to demonstrate its effectiveness in more complex scenarios. These experiments will not only validate the mechanism but also refine it for advanced applications, providing valuable insights for enhancing robot-human interactions.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 14 September 2023; Accepted: 2 August 2024
Published online: 03 September 2024

References

1. Katsuki, F. & Constantinidis, C. Bottom-up and top-down attention: Different processes and overlapping neural systems. *Neuroscientist* **20**, 509–521 (2014).
2. Posner, M. I., Rothbart, M. K. & Ghassemzadeh, H. Developing attention in typical children related to disabilities. In *Handbook of Clinical Neurology*, vol. 173, 215–223 (Elsevier, 2020).
3. Lewkowicz, D. J. & Hansen-Tift, A. M. Infants deploy selective attention to the mouth of a talking face when learning speech. *Proc. Nat. Acad. Sci.* **109**, 1431–1436 (2012).
4. Kuhl, P. K. & Meltzoff, A. N. The bimodal perception of speech in infancy. *Science* **218**, 1138–1141 (1982).
5. Guellaï, B. *et al. Sensus communis*: Some perspectives on the origins of non-synchronous cross-sensory associations. *Front. Psychol.* **10**, 523 (2019).
6. Itti, L. & Baldi, P. F. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems*, 547–554 (2006).
7. Nakajima, J., Kimura, A., Sugimoto, A. & Kashino, K. Visual attention driven by auditory cues. In *International Conference on Multimedia Modeling*, 74–86 (Springer, 2015).
8. Schauerte, B. & Stiefelhagen, R. "wow!" bayesian surprise for salient acoustic event detection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6402–6406 (IEEE, 2013).
9. Bredin, H. & Chollet, G. Audio-visual speech synchrony measure for talking-face identity verification. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, II–233 (IEEE, 2007).

10. Kidron, E., Schechner, Y. Y. & Elad, M. Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 88–95 (IEEE, 2005).
11. Sargin, M. E., Yemez, Y., Erzin, E. & Tekalp, A. M. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Trans. Multimed.* **9**, 1396–1403 (2007).
12. Iyengar, G., Nock, H. J. & Neti, C. Audio-visual synchrony for detection of monologues in video archives. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)*, vol. 5, V–772 (IEEE, 2003).
13. Rolf, M., Hanheide, M. & Rohlfing, K. J. Attention via synchrony: Making use of multimodal cues in social learning. *IEEE Trans. Auton. Ment. Dev.* **1**, 55–67 (2009).
14. Coutrot, A. & Guyader, N. An audiovisual attention model for natural conversation scenes. In *2014 IEEE International Conference on Image Processing (ICIP)*, 1100–1104 (IEEE, 2014).
15. Goldberg, J. & Schoner, G. Understanding the distribution of infant attention: A dynamical systems approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29 (2007).
16. Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259 (1998).
17. Sidaty, N., Larabi, M.-C. & Saadane, A. Toward an audiovisual attention model for multimodal video content. *Neurocomputing* **259**, 94–111 (2017).
18. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).
19. Begum, M. & Karray, F. Visual attention for robotic cognition: A survey. *IEEE Trans. Auton. Ment. Dev.* **3**, 92–105 (2010).
20. Chen, Y. *et al.* Audio matters in visual attention. *IEEE Trans. Circuits Syst. Video Technol.* **24**, 1992–2003 (2014).
21. Hasnain, S. K., Mostafaoui, G. & Gaussier, P. A synchrony-based perspective for partner selection and attentional mechanism in human–robot interaction. *Paladyn* **3**, 156–171 (2012).
22. Heckmann, M. *et al.* An audio-visual attention system for online association learning. In *10th Annual Conference of the International Speech Communication Association*, 2171–2174 (2009).
23. Hori, C. *et al.* End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2352–2356 (IEEE, 2019).
24. Oliva, A., Torralba, A., Castelhano, M. S. & Henderson, J. M. Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing*, vol. 1, I–253 (IEEE, 2003).
25. Quigley, C., Onat, S., Harding, S., Cooke, M. & König, P. Audio-visual integration during overt visual attention. *J. Eye Mov. Res.* **1**, 1–17 (2008).
26. Saalmann, Y. B., Pigarev, I. N. & Vidyasagar, T. R. Neural mechanisms of visual attention: How top-down feedback highlights relevant locations. *Science* **316**, 1612–1615 (2007).
27. Posner, M. I. Orienting of attention. *Q. J. Exp. Psychol.* **32**, 3–25 (1980).
28. Amso, D., Haas, S., Tenenbaum, E., Markant, J. & Sheinkopf, S. J. Bottom-up attention orienting in young children with autism. *J. Autism Dev. Disord.* **44**, 664–673 (2014).
29. Anderson, P. *et al.* Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086 (2018).
30. Smith, L. B., Jones, S. S. & Landau, B. Naming in young children: A dumb attentional mechanism?. *Cognition* **60**, 143–171 (1996).
31. Talsma, D., Senkowski, D., Soto-Faraco, S. & Woldorff, M. G. The multifaceted interplay between attention and multisensory integration. *Trends Cogn. Sci.* **14**, 400–410 (2010).
32. Cangelosi, A. & Schlesinger, M. *Developmental Robotics: From Babies to Robots* (MIT Press, 2015).
33. Cangelosi, A. & Schlesinger, M. From babies to robots: The contribution of developmental robotics to developmental psychology. *Child. Dev. Perspect.* **12**, 183–188 (2018).
34. Pfeifer, R., Lungarella, M. & Iida, F. Self-organization, embodiment, and biologically inspired robotics. *Science* **318**, 1088–1093 (2007).
35. Lewkowicz, D. J. Perception of auditory-visual temporal synchrony in human infants. *J. Exp. Psychol. Hum. Percept. Perform.* **22**, 1094 (1996).
36. Horn, B. K. & Schunck, B. G. Determining optical flow. In *Techniques and Applications of Image Understanding*, vol. 281, 319–331 (International Society for Optics and Photonics, 1981).
37. Rolf, M. Audiovisual attention via Synchrony. Ph.D. thesis, Master's thesis, Bielefeld University (2008).
38. Nock, H. J., Iyengar, G. & Neti, C. Speaker localisation using audio-visual synchrony: An empirical study. In *International Conference on Image and Video Retrieval*, 488–499 (Springer, 2003).
39. Hebb, D. O. The first stage of perception: Growth of the assembly. *Org. Behav.* **4**, 60–78 (1949).
40. Cohen, J. Y. *et al.* Cooperation and competition among frontal eye field neurons during visual target selection. *J. Neurosci.* **30**, 3227–3238 (2010).
41. Diamond, A. Executive functions. *Annu. Rev. Psychol.* **64**, 135–168 (2013).
42. Houdé, O. Inhibition and cognitive development: Object, number, categorization, and reasoning. *Cogn. Dev.* **15**, 63–73 (2000).
43. Curtindale, L. M., Bahrick, L. E., Lickliter, R. & Colombo, J. Effects of multimodal synchrony on infant attention and heart rate during events with social and nonsocial stimuli. *J. Exp. Child Psychol.* **178**, 283–294 (2019).
44. de Villiers Rader, N. & Zukow-Goldring, P. Caregivers' gestures direct infant attention during early word learning the importance of dynamic synchrony. *Lang. Sci.* **34**, 559–568 (2012).
45. Boucenna, S., Gaussier, P., Andry, P. & Hafemeister, L. A robot learns the facial expressions recognition and face/non-face discrimination through an imitation game. *Int. J. Social Robot.* **6**, 633–652 (2014).
46. Boucenna, S., Cohen, D., Meltzoff, A. N., Gaussier, P. & Chetouani, M. Robots learn to recognize individuals from imitative encounters with people and avatars. *Sci. Rep.* **6**, 19908 (2016).
47. Pitti, A., Kuniyoshi, Y., Quoy, M. & Gaussier, P. Modeling the minimal newborn's intersubjective mind: The visuotopic-somatotopic alignment hypothesis in the superior colliculus. *PLoS One* **8**, e69474 (2013).
48. Ellis, C. T., Skalaban, L. J., Yates, T. S. & Turk-Browne, N. B. Attention recruits frontal cortex in human infants. *Proc. Nat. Acad. Sci.* **118**, e2021474118 (2021).
49. Ayneto, A. & Sebastian-Galles, N. The influence of bilingualism on the preference for the mouth region of dynamic faces. *Dev. Sci.* **20**, e12446 (2017).
50. Jones, W., Carr, K. & Klin, A. Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Arch. Gen. Psychiatry* **65**, 946–954 (2008).
51. Gliga, T. & Csibra, G. Seeing the face through the eyes: A developmental perspective on face expertise. *Prog. Brain Res.* **164**, 323–339 (2007).
52. Amso, D., Fitzgerald, M., Davidow, J., Gilhooly, T. & Tottenham, N. Visual exploration strategies and the development of infants' facial emotion discrimination. *Front. Psychol.* **1**, 180 (2010).

Author contributions

S.B. and R.B. conceived the experiments, R.B., R.D. and S.B. conducted the experiments, R.B., S.B., R.D., D.C. and A.P. analysed the results. S.B. and R.B. wrote the manuscript. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to S.B.

Reprints and permissions information is available at www.nature.com/reprints.

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