Towards Crossmodal Learning for Smooth Multimodal Attention Orientation

Frederik Haarslev[1], David Docherty[2], Stefan-Daniel Suvei[1], William K. Juel[1], Leon Bodenhagen[1], Danish Shaikh[2], Norbert Krüger[1], and Poramate Manoonpong[2,3]

[1] SDU Robotics, Maersk Mc-Kinney Moller Institute, University of Southern Denmark
[2] SDU Embodied Systems for Robotics and Learning, Maersk Mc-Kinney Moller Institute, University of Southern Denmark
[3] Bio-inspired Robotics and Neural Engineering Laboratory, School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology
{fh|dado|stdasu|wkj|lebo|danish|norbert|poma}@mmmi.sdu.dk

Abstract. Orienting attention towards another person of interest is a fundamental social behaviour prevalent in human-human interaction and crucial in human-robot interaction. This orientation behaviour is often governed by the received audio-visual stimuli. We present an adaptive neural circuit for multisensory attention orientation that combines auditory and visual directional cues. The circuit learns to integrate sound direction cues, extracted via a model of the peripheral auditory system of lizards, with visual directional cues via deep learning based object detection. We implement the neural circuit on a robot and demonstrate that integrating multisensory information via the circuit generates appropriate motor velocity commands that control the robot's orientation movements. We experimentally validate the adaptive neural circuit for co-located human target and a loudspeaker emitting a fixed tone.

1 Introduction

Orienting spatial attention [15] towards relevant events is a fundamental behaviour in humans. Spatial attention is governed by both top-down, endogenous as well as bottom-up, exogenous mechanisms. Endogenous orientation of spatial attention is driven by the purposeful assignment of neural resources to a relevant and expected spatial target. It is determined by the observer's intent and is a process requiring significant computational resources [13]. For example, when conversing with another person, our mental resources are engaged and our spatial attention is directed towards that person. Exogenous orientation of spatial attention is driven by the sudden appearance of unexpected stimuli in the peripheral sensory space. It is determined by the properties of the stimuli alone and is manifested as an automatic reflexive saccade requiring significantly less computational resources [13]. For example, a loud noise or flash of light in our

sensory periphery directs our attention via orientation of the eyes and/or head towards the spatial location of the event. This occurs even if our attention is focused elsewhere in space, for example when conversing intently with a person. In this article we focus on exogenous spatial attention orientation.

Spatial orientation behaviour is typically driven by the two dominant senses, vision and sound, providing the necessary sensory cues. Orienting towards an audio-visual target outside the visual field must initially engage auditory attention mechanisms. The resultant initial saccade towards the target may be inaccurate since auditory spatial perception is relatively inferior to its visual counterpart. Any error in orientation may then be compensated for by engaging the visual attention mechanisms that bring the target in the centre of the visual field and maintain it there. However, such sequential processing of auditory and visual spatial cues may result in unnecessary saccadic oscillations in more complex tasks. For example, orienting towards an unknown person that unexpectedly calls out our name from a location outside of the visual field. Although audio still initiates the orientation response, both auditory and visual spatial cues (that are also spatially congruent) are needed to generate an optimal orientation response. Processing of such multimodal cues results in smooth and efficient orientation behavior that minimises saccadic oscillations. Audio-visual multisensory cue integration has been studied from the perspective of Bayesian inference [7]. However, Bayesian cue integration implies that *a priori* auditory and visual estimates of spatial location as well as of their relative reliabilities are available. For a robot interacting with a human in a natural setting, the aforementioned *a priori* information cannot always be foreseen and integrated into the robot's programming.

We present an adaptive neural circuit for smooth exogenous spatial orientation. It fuses auditory and visual directional cues via weighted cue integration computed by a single multisensory neuron. The neural circuit adapts sensory cue weights, initially learned offline in simulation, online using bi-directional cross-modal learning via the Input Correlation (ICO) learning algorithm [14]. The proposed cue integration differs from true Bayesian cue integration in that no *a priori* knowledge of sensory cue reliabilities is required to determine the sensory cue weights. The neural circuit is embodied as a high-level adaptive controller for a mobile robot that must localise an audio-visual target by orienting smoothly towards it. We experimentally demonstrate that online adaptation of the sensory cue weights, initially learned offline for a given target location, reduces saccadic oscillations and improves the orientation response for a new target location.

2 Related work

A comprehensive review of multimodal fusion techniques through a number of classifications based on the fusion methodology as well as the level of fusion can be found in [3]. There are a number of techniques reported in the literature that perform audio-visual fusion in the context of speaker tracking. Conventional approaches rely on particle filtering [16,12] as well as Kalman filtering and its

extensions such as decentralized Kalman filters [6] and extended Kalman filters [8]. Other techniques reported in the literature include location-based weighted fusion [11], audio-visual localisation cue association based on Euclidean distance [20], Gaussian mixture models [18] and Bayesian estimation [10]. The goal of the present work is not to improve upon the numerous existing approaches to audio-visual spatial localisation. The majority of these systems either focus only on passive localisation or decouple the computations required for generating the subsequent motor behaviour from the computations performed for localisation. Spatial localisation in humans on the other hand utilises multimodal cues and is tightly coupled to the inevitable action that is subsequently performed, i.e. smoothly orienting towards the target. In human-robot interaction this natural and seemingly ordinary behaviour influences the trustworthiness of the robot [4] and hence the applicability of such a robot to real-world tasks. [2] have experimentally investigated user localisation and spatial orientation via multimodal cues during human-robot interaction. However, they process auditory and visual sensor information sequentially to perform localisation. Furthermore, they decouple localisation from spatial orientation. We, on the other hand, present a neural learning architecture for crossmodal integration that tightly couples audio-visual localisation with smooth exogenous spatial orientation.

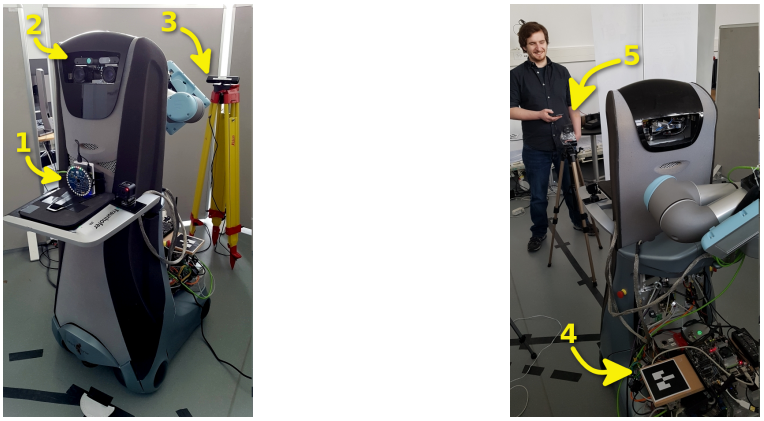
3 Materials and methods

In the following, an overview of the robotic platform, the processing of audio and visual signals and the framework for fusing both signals is provided, as well as the experimental setup.

The robot platform The Care-O-Bot (see Fig. 1) [9] is a research platform, developed to function as a mobile robot assistant that actively supports humans, e.g. in activities of daily living. It is equipped with various sensors and has a modular hardware setup, which makes it applicable for a large variety of tasks. The main components of the robot are: the omni-directional base, an actuated torso, the head containing a Carmine 3D Sensor and a high resolution stereo camera, as well as three laser scanners used for safety and navigation.

Auditory processing The auditory directional cue is extracted by a model of the peripheral auditory system of a lizard [5]. The model maps the minuscule phase differences between the input sound signals into relatively larger differences in the amplitudes of the output signals. Since the phase difference corresponds to the sound direction, the direction can be be formulated as a function of the sound amplitudes:

$$\left|\frac{i_{\mathrm{I}}}{i_{\mathrm{C}}}\right| = 20\left(\log|i_{\mathrm{I}}| - \log|i_{\mathrm{C}}|\right)\mathrm{dB}. \tag{1}$$



(a) Front view of the robot. (b) Back view of the robot.

Fig. 1: The Care-O-Bot platform with key components highlighted: microphone array (1), cameras (2, 3), AR marker (4), and loudspeaker (5).

where $|i_{\mathrm{I}}|$ and $|i_{\mathrm{C}}|$ model the vibration amplitudes of the ipsilateral and contralateral eardrums. The sound direction information in (1) is subsequently normalised to lie within ± 1. Therefore, the auditory directional cue can now be formulated as

$$x_{\mathrm{a}} = \frac{\left|\frac{i_{\mathrm{I}}}{i_{\mathrm{C}}}\right|}{\max\limits_{-\frac{\pi}{2} \geq \theta \leq +\frac{\pi}{2}} \left|\frac{i_{\mathrm{I}}}{i_{\mathrm{C}}}\right|} \quad . \tag{2}$$

where θ is the sound direction. The model is implemented as a 4th-order digital bandpass IIR filter. The auditory direction cue as given by (1) is used as the auditory input x_{a} to the adaptive neural circuit. The peripheral auditory system, its equivalent circuit model and response characteristics, have been reported earlier in detail [19]. The model's frequency response is dependent on the phase differences between the input sound signals, which in turn is dependent on the physical separation between the microphones used to capture the sound signals.

An off-the-shelf multi-microphone array (Matrix Creator[4]) was used to capture the raw sound signals. The microphones were 40 mm apart, resulting in the model's frequency response lying within the range 400 Hz–700 Hz. This range is within the bounds of human speech fundamentals and harmonics (100 Hz to 17 kHz) whilst avoiding the background noise of the robot (approx. 258 Hz) and experimental arena (approx. 20 kHz).

Visual processing For the visual perception of the robot, the convolutional neural network YOLOv2 [17] was applied on 2D images taken with a Carmine

[4] `www.matrix.one/products/creator`

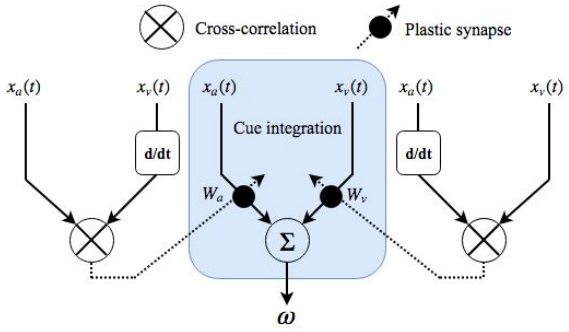


Fig. 2: The adaptive neural circuit. $x_{\mathrm{a}}(t)$ and $x_{\mathrm{v}}(t)$ are respectively the auditory and visual directional cues extracted by the robot that are fused to compute the robot's angular velocity ω. Synaptic weights, w_{v} and w_{a}, respectively scale the directional cues.

sensor. YOLO is an object detection network showing state of the art performance on various object detection benchmarks. It is also significantly faster than other object detection architectures released since 2016. Since the computations are performed on a NVIDIA Jetson TX2[5], the YOLO-tiny variant is used resulting in a framerate of 5 Hz. The network outputs a bounding box for each detection, containing the centre of the box (u,v) and its size. Since only the relative direction of the person is required, only the horizontal position v is used. This is normalised with the image width to produce a number between ± 1.

Crossmodal learning Fig. 2 depicts the adaptive neural circuit for crossmodal integration. A single multisensory neuron computes the angular velocity ω of the robot as the weighted sum of auditory and visual directional cues x_{a} and x_{v} respectively. Audio-visual cue integration is therefore modelled as

$$\omega = w_{\mathrm{v}} x_{\mathrm{v}}(t) + w_{\mathrm{a}} x_{\mathrm{a}}(t) \tag{3}$$

In (3) w_{v} and w_{a} are the synaptic weights that respectively scale the visual and the auditory directional cues. For updating the weights, two learning rules that reflect bi-directional crossmodal integration are defined:

$$\frac{\delta w_{\mathrm{v}}(t)}{\delta t} = \mu x_{\mathrm{v}}(t) \frac{\delta x_{\mathrm{a}}(t)}{\delta t} \qquad \frac{\delta w_{\mathrm{a}}(t)}{\delta t} = \mu x_{\mathrm{a}}(t) \frac{\delta x_{\mathrm{v}}(t)}{\delta t} \tag{4}$$

Both the learning rules employ the same learning rate μ. In either learning rule, the directional cue from one modality is multiplied with the time derivative of the directional cue from the other modality. Therefore, (4) represent cross-correlations between one directional cue and the rate of change of the other.

[5] `www.nvidia.com/en-us/autonomous-machines/embedded-systems-dev-kits-modules/?section=jetsonDevkits`

There are no vision weight updates when either the visual cue becomes zero and/or the auditory cue becomes constant or zero. This mechanism ensures that the weight updates progressively get relatively smaller the closer the target moves to the centre of the FoV and the slower it moves. This allows the weights to stabilise when the robot is pointing directly towards the target. A similar argument can be made for the auditory weight updates. Such bi-directional crossmodal learning allows both the visual and auditory cue weights to stabilise by compensating for errors in the directional cues extracted from either modality.

When the target is outside the FoV the visual cue x_{v} is zero. Therefore, the visual and auditory cue weights w_{v} and w_{a} are not updated and remain fixed at their initial values. The robot's turning behaviour initially depends only on the magnitude of the auditory cue. As the robot keeps turning, the human subject eventually appears within the FoV. Both visual and auditory cues x_{v} and x_{a} then become non-zero. As the robot continues to turn towards the human-loudspeaker target, it comes closer to the centre of the robot's auditory and FoV. Consequently, both the visual and auditory directional cues gradually decrease towards 0. The angular velocity ω, computed by (3), will also gradually decrease as a result. The robot should stop turning when it is aligned with the target.

Experimental setup The task of the robot in the experimental arena (Fig. 3), is to align towards an audio-visual target represented by a human subject (P) co-located with a loudspeaker (S). The angular position of the target relative to the robot's initial orientation is defined as left for $-45°$ and right for $45°$. The initial orientation of the robot in all trials is facing forward, defined as $0°$. The robot must adaptively fuse visual and auditory directional cues to generate appropriate motor velocity commands to orient towards the target. The adaptation comes from learning appropriate sensory cue weights w_{v} and w_{a}, respectively for the visual and auditory signals. The weights are initially learned offline in simulation and then adapted online to smoothen the orientation movements of the robot for targets not encountered previously.

Simulation trials: The sensory cue weights of the neural circuit are first learned offline in simulation, using an instance of the neural circuit. In the simulation the target is placed on the right, meaning that the the weights learned offline represent optimised values for the target located to the right.

The weights w_{a} and w_{v} are randomly initialised to values between 0.01 and 0.05. At each simulation time step in a single trial, two simulated 600 Hz sinusoids, phase-shifted according to sound source location and microphone separation, are input to the ear model. These sinusoids model a loudspeaker emitting a 600 Hz tone from the target position. The normalised output x_{a} of the ear model maps to angular positions $\pm 90°$ relative to the initial orientation. The neural circuit computes the angular velocity using (3) and this orients the robot towards the target. As the target enters the FoV, the normalised visual directional cue x_{v}, between ± 1 is generated. This maps to a FoV of approx. $\pm 29°$ relative to the initial orientation. The weights w_{a} and w_{v} are subsequently updated via the ICO learning rules given by (4).

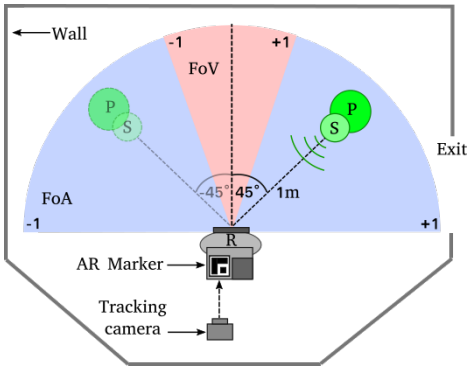


Fig. 3: Experimental setup where a loudspeaker (S) is placed 1 m away from robot (R) at an offset from the centre by ±45° and with a person (P) standing just behind it. The field of view (FoV) is approx. ±29° and the field of audio (FoA) is approx. ±90°.

We quantify the orientation performance in terms of the orientation error. The orientation error is defined as the difference between the robot's orientation after any oscillations have died out and the target's angular position. We determine the average orientation error over a set of 10 trials with randomly initialised, but identical sensory cue weights. We perform this step 30 times to get 30 values for the average orientation error. We then perform an additional trial using, as the initial weights, the initial weights for the set with the lowest average orientation error. The weights learned at the end of this trial ($w_{\mathrm{a}} = 0.027744$, $w_{\mathrm{v}} = 0.034845$) are deemed as the optimised, offline-learned weights.

Real world trials: The target is a human subject co-located with a loudspeaker emitting a 600 Hz tone. The real-world trials use another instance of the neural circuit that can adapt the offline-learned weights further, to generate smooth orientation movements. We perform two sets of trials, one where the target is located to the right and another where the target is located to the left. We perform 20 trials for each target location, where 10 trials are without online learning and 10 trials are with online learning. Therefore, 40 trials are performed in total. In all trials, the neural circuit is initialised with the offline-learned, optimised values for w_{a} and w_{v}.

A PrimeSense 3D sensor in conjunction with the ALVAR [1] software library tracks an AR marker attached to the robot (Fig. 1b). The tracking data is used to determine rotation angle of the robot relative to its initial orientation. The goal configuration, i.e. the robot facing the target and the person being in the center of the FoV, is identified manually and used as ground truth. We quantify the orientation performance of the robot in terms of the orientation error and time taken for any oscillations in the robot's movement to settle. The orientation error is defined as the difference between the robot's orientation after any oscillations have died out and the goal configuration. We define the time taken for the

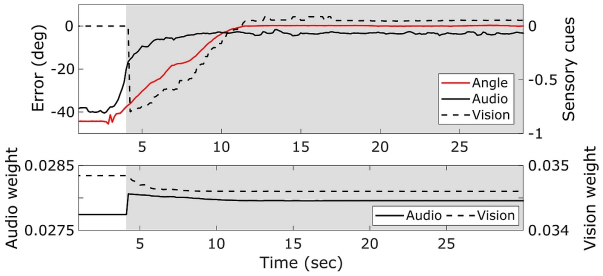
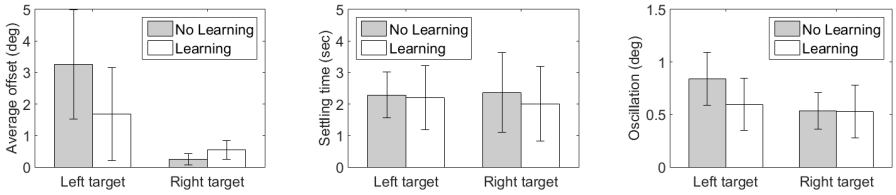


Fig. 4: Recordings from a single trial, with the target located on the left. Top: Auditory (solid black line) and visual (dotted lines) cues; the red line shows the orientation of the robot relative to the target. Bottom: weights for auditory (solid line) and visual (dotted line) cue. Shaded regions indicate the period in which audio-visual fusion occurs.



(a) Average offset. (b) Average settling time. (c) Average oscillations.

Fig. 5: Average results for the turning behaviour with and without learning with error bars indicating the standard deviation.

oscillations in the robot's movements to settle as the oscillation period. It is determined as the time from the first overshoot to when the standard deviation in orientation error reduces to below 0.3°.

4 Results

In this section we present the results from the real-world trials. Fig. 4 shows experimental data from a single trial where the development of the sensory cues, the corresponding weights and the orientation error is visible. It is evident that the orientation error initially decreases relatively slowly, when only the auditory cue is available. Once the visual cue becomes available (i.e. non-zero) the neural circuit fuses the two together to adaptively orient the robot towards the target.

The average performance of the turning behaviour it shown on Fig. 5 for both target configurations with and without learning. Since the offline weights are optimised for a target on the right side, significant improvement cannot be expected on that side. Using the offline weights for orienting to the left without fine-tuning them online results in greater orientation error in general. In this case, using online learning to further fine-tune the weights proves beneficial as it

reduces orientation error significantly. This supports our hypothesis that online fine-tuning of the weights smoothes the orientation movements of the robot for a target not encountered previously.

For assessing the effect of learning a two-tailed t-test with equal variances not assumed has been conducted. For the left side, online learning reduces the offset by 49 % in average and significantly ($p = 0.041$) improves the robot behaviour. For the right side, online learning leads to a marginal increase of the offset ($p = 0.020$).

The oscillations are found to be reduced significantly for the left target ($p = 0.043$) while no difference was observed for the right target. No significant effect has been found for the settling time although the trend for this measure was slightly positive for both targets.

5 Conclusion and future work

We have presented an adaptive neural circuit for multimodal and smooth exogenous spatial attention orientation, in a human-robot interaction scenario. The circuit adaptively fuses auditory and visual directional cues online to orient a mobile robot towards an audio-visual target. We first learned the auditory and visual cue weights offline in simulation for a target located on the right only. We adapted the weights via online learning in real world trials for targets located on both the left and the right of the robot. We determined the orientation error and time taken for possible oscillations in robot's movements to settle. For the target to the left, we observed significant improvement in orientation error with online learning as compared to without online learning. This supports our hypothesis that fine-tuning of the weights via online learning smoothes the orientation movements of the robot for a target not encountered previously.

The smooth spatial orientation behaviour can be subsequently extended to smoothly approach a human subject. Smooth approach can be achieved by extending the adaptive neural circuit to include the depth information. The sound localisation used here can be extended to localise natural human speech by combining multiple ear models with varying sound frequency responses.

Acknowledgement

This research was part of the SMOOTH project (project number 6158-00009B) by Innovation Fund Denmark.

References

1. Alvar 2.0, `http://docs.ros.org/api/ar_track_alvar/html/`
2. Alonso-Martín, F., Gorostiza, J.F., Malfaz, M., Salichs, M.A.: User localization during human-robot interaction. Sensors **12**(7), 9913–9935 (2012)
3. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Systems **16**(6), 345–379 (11 2010)

4. van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Haselager, P.: Do robot performance and behavioral style affect human trust? Int. Journal of Social Robotics **6**(4), 519–531 (11 2014)
5. Christensen-Dalsgaard, J., Manley, G.: Directionality of the Lizard Ear. Journal of Experimental Biology **208**(6), 1209–1217 (2005)
6. D'Arca, E., Robertson, N.M., Hopgood, J.: Person tracking via audio and video fusion. In: 9th IET Data Fusion Target Tracking Conference: Algorithms Applications. pp. 1–6 (2012)
7. David, B., David, A.: Combining visual and auditory information. In: Martinez-Conde, S., Macknik, S., Martinez, L., Alonso, J.M., Tse, P. (eds.) Visual Perception—Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception, Progress in Brain Research, vol. 155, Part B, pp. 243–258. Elsevier (2006)
8. Gehrig, T., Nickel, K., Ekenel, H.K., Klee, U., McDonough, J.: Kalman filters for audio-video source localization. In: IEEE Works. on Applications of Signal Processing to Audio and Acoustics. pp. 118–121 (2005)
9. Graf, B., Reiser, U., Hägele, M., Mauz, K., Klein, P.: Robotic home assistant care-o-bot 3 - product vision and innovation platform. In: IEEE Works. on Advanced Robotics and its Social Impacts (2009)
10. Hoseinnezhad, R., Vo, B.N., Vo, B.T., Suter, D.: Bayesian integration of audio and visual information for multi-target tracking using a CB-member filter. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing. pp. 2300–2303 (2011)
11. Kheradiya, J., C, S.R., Hegde, R.: Active Speaker Detection using audio-visual sensor array. In: IEEE Int. Symposium on Signal Processing and Information Technology. pp. 480–484 (2014)
12. Kiliç, V., Barnard, M., Wang, W., Kittler, J.: Audio Assisted Robust Visual Tracking With Adaptive Particle Filtering. IEEE Trans. on Multimedia **17**(2), 186–200 (2015)
13. Mayer, A.R., Dorflinger, J.M., Rao, S.M., Seidenberg, M.: Neural networks underlying endogenous and exogenous visual–spatial orienting. NeuroImage **23**(2), 534–541 (2004)
14. Porr, B., Wörgötter, F.: Strongly improved stability and faster convergence of temporal sequence learning by utilising input correlations only. Neural Computation **18**(6), 1380–1412 (2006)
15. Posner, M.I.: Orienting of attention. Quarterly Journal of Experimental Psychology **32**(1), 3–25 (1980)
16. Qian, X., Brutti, A., Omologo, M., Cavallaro, A.: 3D audio-visual speaker tracking with an adaptive particle filter. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing. pp. 2896–2900 (2017)
17. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. arXiv preprint arXiv:1612.08242 (2016)
18. Sanchez-Riera, J., Alameda-Pineda, X., Wienke, J., Deleforge, A., Arias, S., Čech, J., Wrede, S., Horaud, R.: Online multimodal speaker detection for humanoid robots. In: 12th IEEE-RAS Int. Conf. on Humanoid Robots. pp. 126–133 (2012)
19. Shaikh, D., Hallam, J., Christensen-Dalsgaard, J.: From "ear" to there: a review of biorobotic models of auditory processing in lizards. Biological Cybernetics **110**(4), 303–317 (2016)
20. Talantzis, F., Pnevmatikakis, A., Constantinides, A.G.: Audio-Visual Active Speaker Tracking in Cluttered Indoors Environments ^{ast}. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **39**(1), 7–15 (2009)