Unimodal Learning Enhances Crossmodal Learning in Robotic Audio-Visual Tracking

Danish Shaikh[1], Leon Bodenhagen[2] and Poramate Manoonpong[1]

***Abstract*— Crossmodal sensory integration is a fundamental feature of the brain that aids in forming an coherent and unified representation of observed events in the world. Spatiotemporally correlated sensory stimuli brought about by rich sensorimotor experiences drive the development of crossmodal integration, with neuroplasticity as its underlying mechanism. Bayesian causal inference framework dictates a weighted combination of stimulus estimates to achieve optimal crossmodal integration, but assumes knowledge of the underlying stimulus noise distributions. We present a Hebbian-like correlation learning-based model that continuously adapts crossmodal combinations in response to dynamic changes in noisy sensory stimuli but does not require a priori knowledge of sensory noise. The model correlates sensory cues within a single modality as well as across modalities to independently update modality-specific neural weights. This model is instantiated as a neural circuit that continuously learns the best possible weights required for a weighted combination of noisy low-level auditory and visual spatial target direction cues. The combined sensory information is directly mapped to wheel velocities that orient a non-holonomic robotic agent towards a moving audio-visual target. Simulation results demonstrate that unimodal learning enhances crossmodal learning and improves both the overall accuracy and precision of multisensory orientation response.**

I. INTRODUCTION

We constantly perceive and localise moving objects in our environment using crossmodal cues. Vision and audition are the primary modalities involved in spatial localisation tasks. The integration of auditory and visual cues forms a unified percept of the motion of an audio-visual target. It has been shown that visual spatial acuity is biased towards the frontal space than at the periphery, while auditory spatial acuity is biased towards the periphery than in the frontal space [4]. This implies that during bimodal localisation auditory and visual biases either dominate the other or are combined together via a neural weighting scheme. Indeed, fMRI studies suggest that both visual and auditory neural substrates share resources [5]. In the *superior colliculus*, a central region for multisensory integration in the midbrain, visual and auditory afferent signals converge on to many multisensory neurons [6]. When two different sensory stimuli are present at close spatial proximity, the neuron exhibits crossmodal enhancement—the neuron's response is significantly greater than that evoked by the most effective of the two unimodal inputs individually [7]. Enhanced orientation behaviour, characterised by reduced reaction time and an increase in probability of correct orientation to a visual target by a spatially coincident auditory stimulus, has been reported in the context of responses of multisensory SC neurons [8], [9]. Audio-visual multisensory integration has been investigated within the Bayesian framework [10]. We have previously reported on audio-visual guidance of orientation behaviour in robotic localisation of a static target [11].

Bayesian models (for a review see [12]) at the single-neuron as well as population level are used to model multisensory integration and predict optimal cue combination at the behavioural level but do not shed light on the underlying mechanisms. These models typically assume that neuronal responses encode likelihood functions with multivariate Gaussian or Poisson distributions. Predictions made by Bayesian models corroborate results of psychophysical bimodal localisation experiments that demonstrate domination of the more reliable cue over the other [12]. In most of these models the neural weight assigned to each modality is typically assumed to be fixed.Neurophysiological evidence indicates that multisensory neuron responses demonstrate reliability-based cue weighting [13], [14], i.e. the weights increase or decrease with relative cue reliability. This implies that learning is a core process in multisensory integration.

Whether learning in multisensory integration is unimodal, crossmodal or both is not well known, but intuition suggests it should be a combination of both. Given that sensory cue reliabilities can independently change over time, crossmodal-only learning, where one unimodal cue modifies the neural weight of another unimodal cue, may cause incorrect weight updates. This is relevant in audio-visual tracking tasks. If at a given instant in time an auditory cue that is relatively less reliable in the frontal space causes the neural weight for the visual cue to be updated, the visual estimate of spatial location in the frontal space may be reduced, leading to a relatively worse multisensory estimate of spatial location after weighted cue combination. Therefore, one can assume that the cue having greater reliability should influence the neural weight of the cue having relatively smaller reliability and not the other way round. This however does not necessarily guarantee optimal or near-optimal cue integration. Unimodal learning for a given modality may also simultaneously influence the associated neural weight independently to generate near-optimal predictions of sensory cues of that modality. This may further improve multisensory responses.

*This research was supported with a grant for the SMOOTH project (project number 6158-00009B) by Innovation Fund Denmark

[1]Danish Shaikh and Poramate Manoonpong are with the Embodied AI and Neurorobotics Laboratory, Centre for BioRobotics, Maersk Mc-Kinney Moeller Institute, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark danish@mmmi.sdu.dk, poma@mmmi.sdu.dk

[2]Leon Bodenhagen is with SDU Robotics, The Maersk Mc-Kinney Moeller Institute, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark lebo@mmmi.sdu.dk

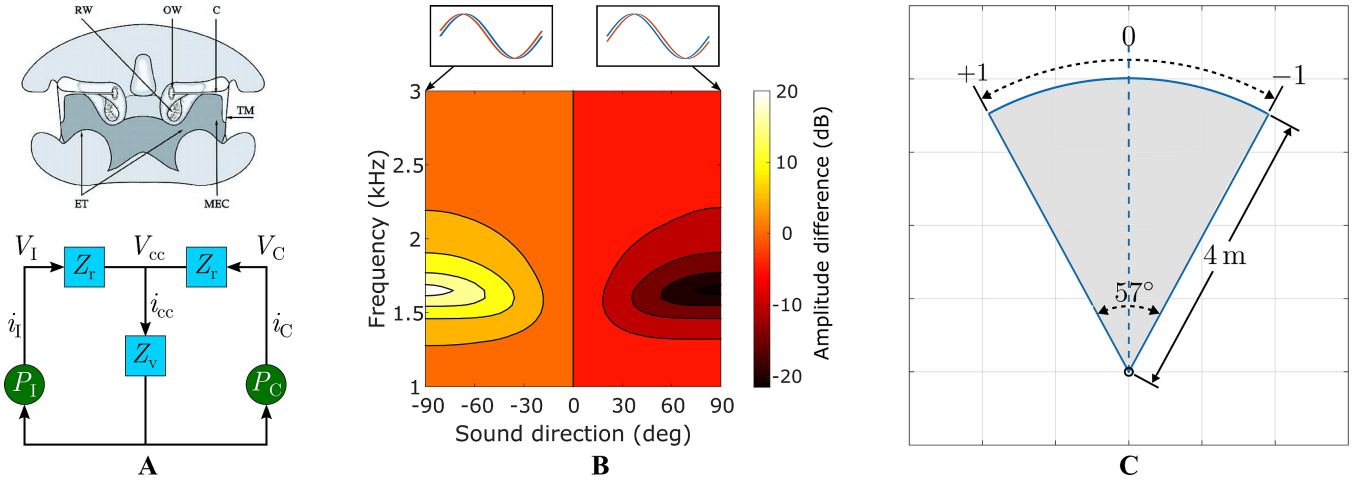

Fig. 1. Auditory and visual directional cues. **A.** Cross-section (top) of the *Sceloporus* lizard peripheral auditory system (borrowed from [1]). Sound waves of wavelengths 340-85 mm (corresponding to frequencies of 1-4 kHz) diffract around the animal's head due to its small size (13 mm) and arrive at the ears (TM) with a tiny phase difference (corresponding to a μ-sec interaural time difference or ITD). Air-filled Eustachian tubes (ET) connecting the two ears allow sound to travel between the ears. These sound direction-dependent phase differences (proportional to time differences) are translated into relatively larger sound direction-dependent differences in perceived sound amplitude at either ear. Electrical (bottom) model of the peripheral auditory system (based on [2] and redrawn from [3]). Voltages V_I and V_C respectively model sound pressures at the ipsilateral (closer to the sound source) and contralateral (further away from the sound source) ears. Impedances Z_r and Z_v together model the total acoustic filtering due to the eardrums, ET and the mouth cavity. The resultant sound pressure in this cavity due to interaction of sound waves from either side is modelled by voltage V_{cc}. Current i_{cc} models sound wave propagation inside the mouth cavity. Currents i_I and i_C model the overall eardrum vibrations and represent sound direction information. **B.** Target auditory direction cue represented as a binaural subtraction of currents i_I and i_C given by (1). **C.** Encoding of target direction cue in the visual field. The visual direction cue varies between -1 when the target is located at the extreme right and +1 when it is located at the extreme left.

We test the hypothesis that unimodal learning enhances crossmodal learning and improves both the accuracy and precision of multisensory responses. We present a simple neural circuit for audio-visual cue integration that directly computes the orientation response, as the weighted sum of auditory and visual directional cues, of a non-holonomic mobile robot that is tasked with dynamically tracking a moving audio-visual target. Three independent Hebbian-like correlation-based learning mechanisms, the first two being unimodal while the third being crossmodal, concurrently update the cue weights. We compare tracking performance in five independent trials.

II. MATERIALS AND METHODS

The auditory directional cue is extracted by the lizard peripheral auditory system (Fig. 1**A**). Sound direction information (Fig. 1**B**) extracted by the peripheral auditory model is formulated as

$$\left|\frac{i_I}{i_C}\right| = \left|\frac{G_I \cdot V_I + G_C \cdot V_C}{G_C \cdot V_I + G_I \cdot V_C}\right| \equiv 20\,(\log|i_I| - \log|i_C|)\,\text{dB}. \quad (1)$$

G_I and G_C respectively are sound frequency-specific (1-2.2 kHz) ipsilateral and contralateral gains. These are experimentally determined by laser vibrometry [1] measurements of eardrum vibrations. The gain terms are implemented as 4th-order digital bandpass filters with infinite impulse response. The peripheral auditory system, its equivalent model and response characteristics has been reported earlier [15] in detail and summarised here for the sake of clarity.

The visual directional cue is modelled as the location of the target inside the visual field (Fig. 1**C**), which is 4 m deep with a 57° horizontal angle, chosen to be the same as a Microsoft Kinect V1 camera sensor. The visual signal x_v is zero when the target is outside the field-of-view, and geometrically calculated to lie within the range $[-1,+1]$ when the target is inside the field-of-view.

The audio-visual tracking task is defined as follows. A simulated robotic agent with rotational freedom but not translational freedom must track a simulated, moving audio-visual target by rotating on-the-spot (Fig. 2**C**). The target moves in a straight line from the right side of the robot to the left with a randomly varying linear velocity between 0-10 m/time step for a random number of time steps between 5-10. This simulates intermittent movements prevalent in real world scenarios. The target emits a 2.2 kHz sinusoidal tone with white Gaussian noise with a signal-to-noise ratio of 20 dB added to it to simulate a noisy cue with relatively low reliability. The emission is intermittent with a random duty cycle, i.e. the emission is on for a random number of simulation time steps between 10–15 and off for a random number of simulation time steps between 5–10. This simulates real-world situations where auditory events are non-continuous, exhibit greater noise and therefore relatively low reliability. The visual signal however is continuous and clean, simulating a cue with a relatively high reliability.

The robotic agent is modelled as a non-holonomic differential drive robot with two wheels (Fig. 2**A**). The wheels are separated by a distance $l = 16$ cm. The robot has two simulated acoustic sensors that functionally mimic microphones to capture the auditory signal emitted by the target. The sensors are separated by 13 mm because the peripheral auditory model parameters have been derived for a 13 mm ear

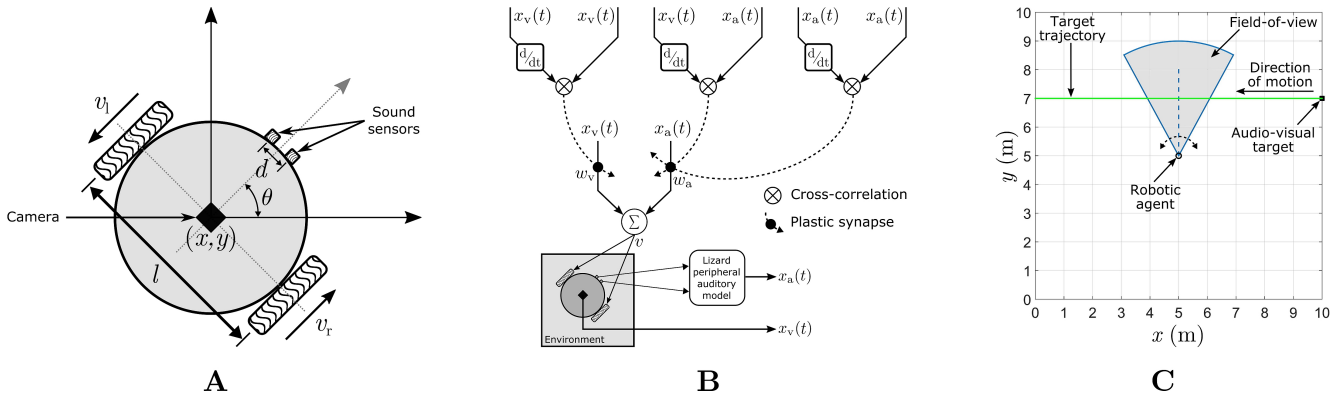

Fig. 2. Experimental setup. **A.** Non-holonomic robot kinematics. **B.** The neural circuit embodied as a robotic agent. **C.** Experimental arena.

separation. The sensor separation must match this to maintain a match between the ITD cues to which the peripheral auditory model is tuned and the actual ITD cues. The lizard peripheral auditory model extracts the auditory direction cue $x_{\mathrm{a}} = \left|\frac{i_{\mathrm{I}}}{i_{\mathrm{C}}}\right|$ from the auditory signal. A virtual visual sensor located at the robot's centre, functionally mimicking a Microsoft Kinect V1 camera, extracts the visual direction cue x_{v} as described earlier. x_{v} and x_{t} are fed to the neural circuit (Fig. 2**B**). The forward kinematic model for differential drive mobile robots [16] as given by (2) is used to determine the pose $[x, y, \theta]$ of the robot, where (x, y) are the two-dimensional coordinates and θ is the orientation.

$$\begin{bmatrix} x \\ y \\ \theta \end{bmatrix} = \begin{bmatrix} cos(\omega\delta t) & -sin(\omega\delta t) & 0 \\ sin(\omega\delta t) & cos(\omega\delta t) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} Dsin(\theta) \\ -Dcos(\theta) \\ \theta \end{bmatrix} + \begin{bmatrix} x - Dsin(\theta) \\ y + Dcos(\theta) \\ \omega\delta \end{bmatrix} \tag{2}$$

where, angular velocity $\omega = \dfrac{(v_{\mathrm{r}} - v_{\mathrm{l}})}{l}$, and distance D from instantaneous center of curvature $= \dfrac{l}{2}\dfrac{(v_{\mathrm{r}} + v_{\mathrm{l}})}{(v_{\mathrm{r}} - v_{\mathrm{l}})}$

At each simulation time step t, a single multisensory neuron computes the common motor velocity $|v| = |v_{\mathrm{l}}| = |v_{\mathrm{r}}|$ of robot as the weighted sum of auditory and visual directional cues x_{a} and x_{v} respectively. The signs for v_{l} and v_{r} are pre-assigned according the direction in which the robot must turn. Thus multisensory integration is modelled as $v = w_{\mathrm{v}} x_{\mathrm{v}}(t) + w_{\mathrm{a}} x_{\mathrm{a}}(t)$. When the visual cue is absent ($x_{\mathrm{v}}(t) = 0$), i.e. when the target is outside the visual field the auditory cue weight w_{a} is independently updated via the unimodal learning rule $\delta w_{\mathrm{a}} = \mu x_{\mathrm{a}}(t)\frac{\delta x_{\mathrm{a}}(t)}{\delta t}$. When the auditory cue is absent ($x_{\mathrm{a}}(t) = 0$), i.e. when the target is not emitting the tone the visual cue weight w_{v} is independently updated via another unimodal learning rule $\delta w_{\mathrm{v}} = \mu x_{\mathrm{v}}(t)\frac{\delta x_{\mathrm{v}}(t)}{\delta t}$. When both auditory and visual direction cues are present, which only occurs when the target is within the visual field and is emitting the tone, the auditory cue weight is independently updated by a third crossmodal learning rule $\delta w_{\mathrm{a}} = \mu x_{\mathrm{a}}(t)\frac{\delta x_{\mathrm{v}}(t)}{\delta t}$. Thus the weight associated with the relatively less reliable auditory cue is also updated according to the correlation between the auditory cue and the relatively more reliable visual cue. Since these three learning mechanisms are independent of each other, the auditory cue weight w_{a} is simultaneously updated via both unimodal and crossmodal learning, while the visual cue weight w_{v} is updated only via unimodal learning because it is assumed to be a clean signal with relatively greater reliability. As discussed earlier, updating the weight of a high reliability cue via a relatively low reliability cue may cause the cue integration result to be degraded. Five independent trials are performed, each with the learning rate μ set to 0.5 and initial values of the weights randomly chosen as $w_{\mathrm{v}} = 0.01$ and $w_{\mathrm{a}} = 0.07$. In all trials the robot is initially pointing straight ahead. Tracking performance is quantified as the relative deviation of the robot's orientation from the target's angular position.

III. RESULTS AND DISCUSSION

For crossmodal-only learning (Fig. 3), the orientation error drops relatively slowly, leftmost graphs) since the visual and auditory cue weights are not updated as long as the target is outside the visual field. When both unimodal and crossmodal learning are allowed (Fig. 3), rightmost graphs), the error drops relatively faster due to unimodal learning outside the visual field. When the target is inside the visual field (Fig. 3, shaded regions), the overall orientation error is visibly greater for crossmodal-only learning as compared to when both unimodal and crossmodal learning are allowed. Both absolute mean and standard deviation values (Fig. 4), calculated over the time period in which audio-visual cue integration occurs (Fig. 3, shaded regions), are relatively lower when both unimodal and crossmodal learning occur concurrently as compared to when crossmodal-only learning occurs. This is in agreement with our hypothesis that unimodal learning enhances crossmodal learning and improves both the accuracy and precision of multisensory responses.

IV. CONCLUSIONS AND FUTURE DIRECTIONS

We presented a neural circuit for multisensory integration that combines unimodal and crossmodal learning. We

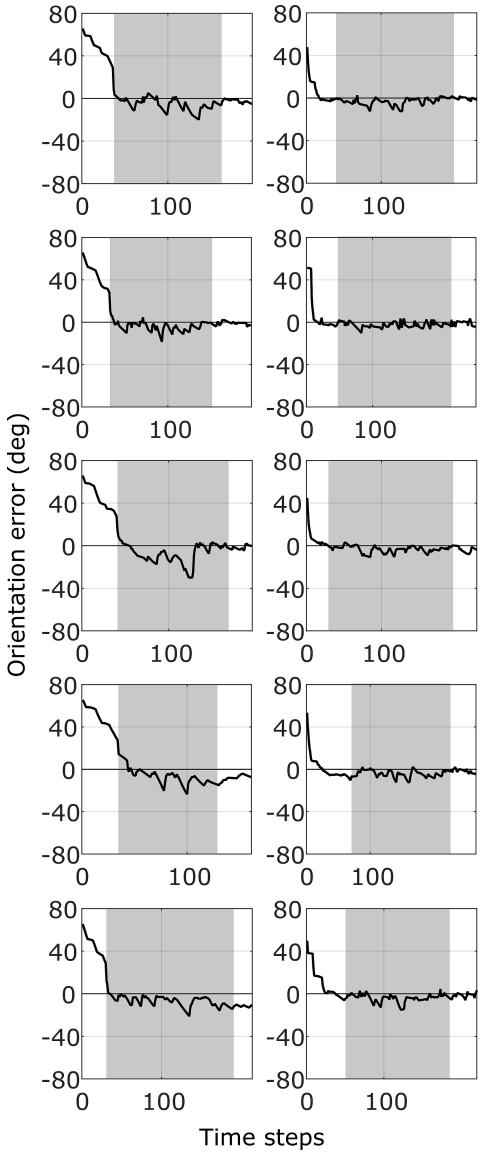

Fig. 3. Orientation error over time. Trials are ordered from top to bottom. For each trial, the left panel corresponds to crossmodal-only learning, while the right panel corresponds to combined unimodal and crossmodal learning. The shaded regions indicate concurrent unimodal and crossmodal learning.

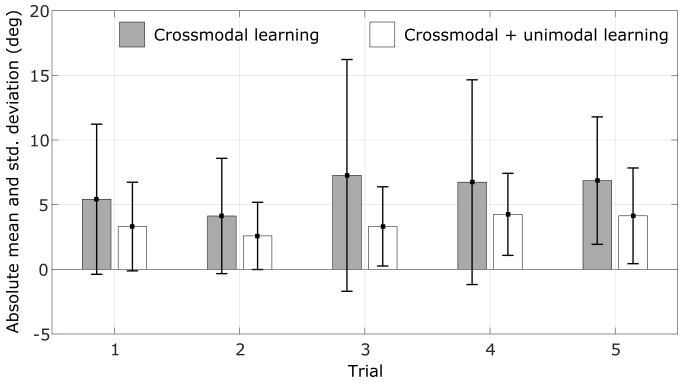

Fig. 4. Absolute mean and standard deviation of orientation error during concurrent unimodal and crossmodal learning.

demonstrated in a simulated audio-visual tracking task that unimodal learning enhances crossmodal learning and improves both the accuracy and precision of orientation responses of a robotic agent. However, neurophysiological evidence for such interplay between unimodal and crossmodal learning is lacking and must be investigated. Furthermore, detailed analysis of the model's behaviour along with greater number of trials is necessary.

REFERENCES

[1] J. Christensen-Dalsgaard and G. Manley, "Directionality of the Lizard Ear," *Journal of Experimental Biology*, vol. 208, no. 6, pp. 1209–1217, 2005.

[2] N. Fletcher, *Acoustic Systems in Biology*. Oxford University Press, USA, 1992.

[3] L. Zhang, "Modelling Directional Hearing in Lizards," Ph.D. dissertation, Maersk Mc-Kinney Moller Institute, Faculty of Engineering, University of Southern Denmark, 2009.

[4] B. Odegaard, D. Wozny, and L. Shams, "Biases in visual, auditory, and audiovisual perception of space," *PLOS Computational Biology*, vol. 11, no. 12, pp. 1–23, 12 2015. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1004649

[5] D. Smith, B. Davis, K. Niu, E. Healy, L. Bonilha, J. Fridriksson, P. Morgan, and C. Rorden, "Spatial Attention Evokes Similar Activation Patterns for Visual and Auditory Stimuli," *Journal of Cognitive Neuroscience*, vol. 22, no. 2, pp. 347–361, 2010.

[6] B. Stein and M. Meredith, *The Merging of the Senses*, ser. A Bradford book. MIT Press, 1993. [Online]. Available: https://books.google.dk/books?id=uCV9QgAACAAJ

[7] M. Wallace, M. Meredith, and B. Stein, "Multisensory Integration in the Superior Colliculus of the Alert Cat," *Journal of Neurophysiology*, vol. 80, no. 2, pp. 1006–1010, 1998. [Online]. Available: http://jn.physiology.org/content/80/2/1006

[8] E. Schröger and A. Widmann, "Speeded responses to audiovisual signal changes result from bimodal integration," *Psychophysiology*, vol. 35, no. 6, pp. 755–759, 1998. [Online]. Available: http://dx.doi.org/10.1111/1469-8986.3560755

[9] B. Stein, M. Meredith, W. Huneycutt, and M. L., "Behavioral Indices of Multisensory Integration: Orientation to Visual Cues is Affected by Auditory Stimuli," *Journal of Cognitive Neuroscience*, vol. 1, no. 1, pp. 12–24, 1989, pMID: 23968407. [Online]. Available: http://dx.doi.org/10.1162/jocn.1989.1.1.12

[10] B. David and A. David, "Chapter 14 combining visual and auditory information," in *Visual Perception—Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception*, ser. Progress in Brain Research, S. Martinez-Conde, S. Macknik, L. Martinez, J.-M. Alonso, and P. Tse, Eds. Elsevier, 2006, vol. 155, Part B, pp. 243–258.

[11] D. Shaikh, P. Manoonpong, G. Tuxworth, and L. Bodenhagen, *Multisensory guidance of goal-oriented behaviour of legged robots*. CLAWAR Association Ltd., 2017, in press.

[12] M. Ursino, C. Cuppini, and E. Magosso, "Neurocomputational approaches to modelling multisensory integration in the brain: A review," *Neural Networks*, vol. 60, pp. 141–165, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608014001981

[13] M. Morgan, G. DeAngelis, and D. Angelaki, "Multisensory Integration in Macaque Visual Cortex Depends on Cue Reliability," *Neuron*, vol. 59, no. 4, pp. 662–673, 2017/08/13 2008. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2008.06.024

[14] C. Fetsch, A. Pouget, G. DeAngelis, and D. Angelaki, "Neural correlates of reliability-based cue weighting during multisensory integration," *Nat Neurosci*, vol. 15, no. 1, pp. 146–154, Jan 2012. [Online]. Available: http://dx.doi.org/10.1038/nn.2983

[15] D. Shaikh, "Exploring a Robotic Model of the Lizard Peripheral Auditory System," Ph.D. dissertation, University of Southern Denmark, 2012.

[16] G. Dudek and M. Jenkin, *Computational Principles of Mobile Robotics*, 2nd ed. New York, NY, USA: Cambridge University Press, 2010.