Feed-forward selection of cerebellar models for calibration of robot sound source localization

Baxendale, M. D.[1,3][0000−0003−4581−4844]*, Nibouche, M.[1][0000−0003−0150−8087], Secco E. L.[3][0000−0002−3269−6749], Pipe, A. G.[2][0000−0002−8404−294X], and Pearson, M. J.[2][0000−0002−8642−4845]

[1] University of the West of England, Bristol, UK
mark2.baxendale@live.uwe.ac.uk, mokhtar.nibouche@uwe.ac.uk
[2] Bristol Robotics Laboratory, Bristol, UK {martin.pearson,tony.pipe}@brl.ac.uk
[3] Liverpool Hope University, Liverpool, UK {mark.baxendale,seccoe}@hope.ac.uk

Abstract. We present a *responsibility predictor*, based on the adaptive filter model of the cerebellum, to provide feed-forward selection of cerebellar calibration models for robot Sound Source Localization (SSL), based on audio features extracted from the received audio stream. In previous work we described a system that selects the models based on sensory feedback, however, a drawback of that system is that it is only able to select a set of calibrators *a-posteriori*, after action (e.g. orienting a camera toward the sound source after a position estimate is made). The responsibility predictor improved the system performance compared to that without responsibility prediction. We show that a trained responsibility predictor is able to use contextual signals in the absence of ground truth to successfully select models with a performance approaching that of a system with full access to the ground truth through sensory feedback.

Keywords: robot audition · responsibility prediction · multiple models · cerebellum · adaptive filter.

1 Introduction

Vision is often used as the primary sensory modality in mobile robots, however, there are situations where vision can become impaired or completely unavailable such as in the aftermath of a disaster. A robot attempting to locate a person in distress, for example, may be unable to do so using vision alone due to airborne particles or collapsed infrastructure. In these situations, audio localization could be used as it does not rely on a direct field of view to the source. Robot audition is a relatively recent field [1] which includes the field of Sound Source Localization (SSL)- the identification of the location of sounds in a robot's environment (azimuth, elevation and distance to source). When operating in challenging acoustic environments such as disaster sites, errors will inevitably occur in the SSL estimation due to reflection, distortion and attenuation of the sound source. Moreover, as the robot navigates through multiple environments the nature of these errors in the SSL estimate will vary according to the acoustic characteristics of those environments. In previous work, we proposed a multiple adaptive filter approach inspired by a cerebellar micro-circuit, to learn to calibrate the output of a SSL system operating in multiple acoustic environments [2]. Subsequently, the system was demonstrated selecting the best calibrator (or set of calibrators) for the robot's current environment, including novel environments through a process of responsibility estimation based on model prediction error. The prediction error is determined through comparison of a model's estimate of the sound source location with the ground truth sound source location. Although that work demonstrated an improvement in SSL, including in comparison to Generalized Cross-Correlation with Phase Transform (GCC-PHAT), a limitation is that it relied on establishing the ground truth location by orienting a camera toward the sound source in order to derive the model prediction errors. This can not always be assumed to be the case for a robot operating in an unstructured disaster site. To make the system more robust we introduce a means of pre-selecting the calibrators before the ground truth becomes available, using features extracted from the audio stream itself. We call this new component the *Responsibility Predictor* (RP) after the Modular Selection And Identification for Control (MOSAIC) framework, which uses the same cerebellar inspired adaptive filter model architecture to learn the contextual cues of the environment [5].

The paper is composed as follows: Section 2 provides background of the biological inspiration and computational implementation of the cerebellar inspired multiple model adaptive filter approach to calibrate a simple SSL algorithm. This is followed in section 3 by a description of the proposed extension to this sensory learning architecture through inclusion of the RP. The experimental apparatus and data capture protocol designed to test the RP are described in the methods section 4, followed by results and discussion of the influence on performance in section 5.

2 Background

2.1 Cerebellar role in binaural sound source localization

Until recently, the cerebellum was considered to mainly be involved in motor control, but there is increasing evidence that it plays a role in non-motor functions, and especially in perceptual processes [8]. The role of the cerebellum in auditory processing in particular was recognised several decades ago [9], but only recently has this aspect of cerebellar function received much attention. Work in this area has mainly focused on speech perception and production; until now there has been very little research on the role of the cerebellum in SSL. A review of binaural SSL in robotics is given in [10]. SSL systems are typically setup in a single, controlled acoustic environment [11, 12], whereas challenging environments such as those described in section 1 can introduce SSL errors, which may depend non-linearly on the azimuth position of the sound source due to complex and unpredictable environmental acoustics. The requirement for non-linear learning was the motivation for applying a computational model of the cerebellum to this problem.

2.2 Adaptive filter model of the cerebellum

The adaptive filter model of the cerebellum was proposed by Fujita [13] and emphasizes the resemblance of the cerebellar microcircuit to an adaptive filter [14], figure 1. The model is characterised by a rich set of inputs/basis filters analogous to the large number of granule cells and Golgi cells in cerebellar cortex, contributing to the power of the adaptive filter function by providing a massive signal analysis capability. This allows a large number of inputs to the model, analogous to the mossy fibres in cerebellum, from very diverse areas of the brain and sensory systems. The parallel fibre signals are synthesised at the Purkinje cell according to the parallel fibre-Purkinje cell synaptic weights, analogous to the summing junction of the adaptive filter.

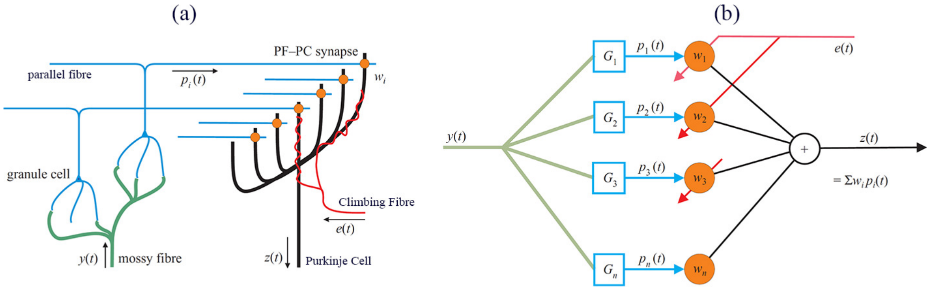

Fig. 1. Adaptive filter model of the cerebellum. (a) Cerebellar microcircuit. (b) Adaptive filter equivalent. Adapted with permission of Royal Society, from [3].

2.3 Multiple internal models and the assignment of responsibility

It has long been hypothesized that the brain possesses internal models to allow the prediction of how the world will respond to actions made, and that the cerebellum is a strong candidate for the site of such models [4]. There are many examples where a single internal model would not be able to capture the range of contexts encountered in real world situations. There have been a number of computational frameworks proposed, including MOSAIC [5] and Hierarchical Attentive Multiple Models for Execution and Recognition of actions (HAMMER) [6], contending that the central nervous system makes use of multiple modules (containing models), each specialised for a specific context. Both frameworks competitively select action, with HAMMER focusing on robot imitation [7]. The MOSAIC framework was used as an inspiration in this study as it has a number of advantages, including proportional combination of module outputs along with both prior and posterior contribution to the production of module probabilities (known in MOSAIC as *responsibilities*), with HAMMER lacking MOSAIC's RP, which underpins the current work.

MOSAIC was developed in the context of motor control, and consists of an array of modules each of which could have influence or control in a particular context. For example the task of lifting an object where objects having differing weights would each represent a different context, requiring a unique force profile. A key problem is that it may not be clear which module would be appropriate until the lifting action has commenced. In MOSAIC, each module consists of a paired forward and inverse model along with a responsibility predictor. There is a single *Responsibility Estimator* (RE) that operates across all modules. Each module receives an efference copy of the motor command, and, within each context, each predicts the state of the system under control as a result of the motor command, assuming it is in the context in which it learnt. After action has commenced or taken place, the ground truth state of the system under control is determined through sensory feedback (such as vision or proprioception), and a prediction error determined for each module. The set of errors is input to the RE, which generates a set of responsibilities using a soft-max function. The responsibilities are used to produce a weighted sum of the control outputs from the modules to form an overall motor command.

This mixing of module outputs in proportion to their responsibilities allows the system to adapt to novel contexts whose characteristics fall between those of contexts in which the modules have learnt. However, the selection of modules is only able to take place when sensory feedback becomes available. The MOSAIC framework includes a *Responsibility Predictor* (RP) that uses contextual signals to predict the responsibility of its module, before sensory feedback is available. By combining the outputs of the RPs with the outputs of the RE, an interplay takes place between the RP and RE. The RP can make an early prediction of the responsibility, potentially reducing performance error when the context changes but the RE outputs have not yet updated in response to sensory feedback. The example used in [15] is that lifting a transparent bottle allows the brain to make a prediction through vision and select appropriate modules for light or heavy objects.

It might seem that an RP that learns to predict its module's responsibility from contextual signals renders the RE redundant. However, the example also points out that an opaque carton would make this impossible, and the brain would need to select models based on prediction error after action has taken place, for which the RE would still be required. Also, the RP could mis-classify the context and select a set of modules that is inappropriate. In [16], Haruno et al. simulated an RP error and showed that in the next time step of the simulation, the RE corrected for the error introduced by the RP once the ground truth had become available through sensory feedback.

2.4 Cerebellar calibration of SSL using Multiple models

The overall system, including the RP developed in this study, is shown in figure 2. It consists of multiple adaptive filter models of the cerebellum each of which learn the SSL error at different azimuths in a given acoustic environment, or context, and adds a compensatory shift to the SSL output. In this study

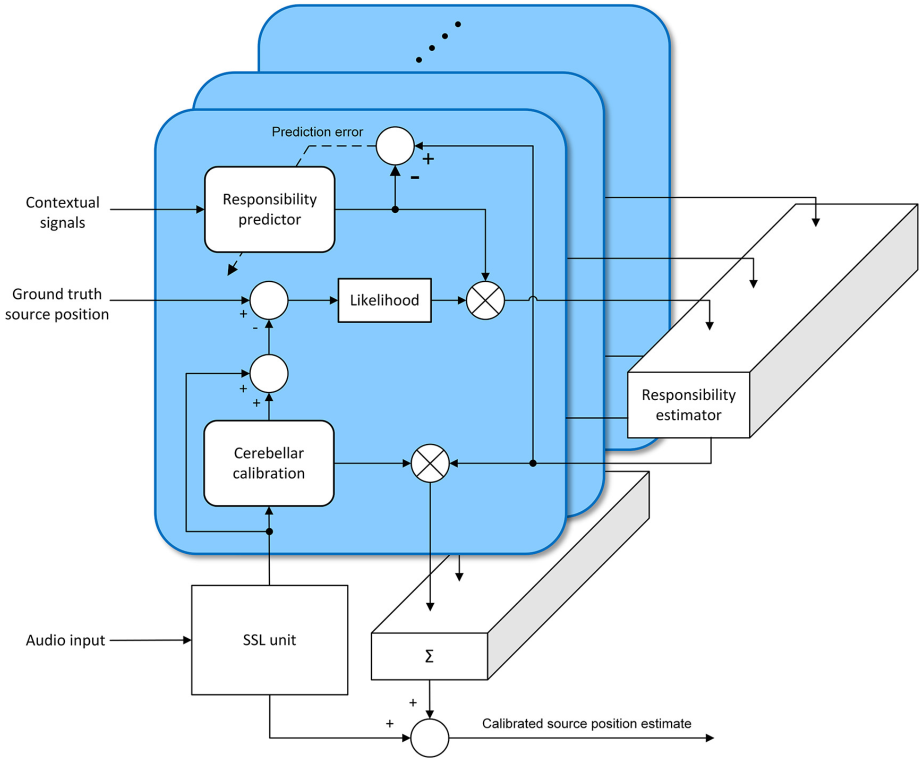

Fig. 2. Responsibility predictor as part of the overall multiple models calibration system.

only azimuth is considered. The cerebellar calibration model was adapted from previous work that calibrated the whisker map of a robot [17]. The amount of calibration required will depend on the azimuth estimate, as different locations within an environment may experience different degrees of error, and is coded in the parallel fibre-Purkinje cell weights through training. A means is required to select the appropriate model for the robot's current environment, and a candidate framework is the MOSAIC described in section 2.3. On receiving audio input the SSL unit makes an estimate of the azimuth position of the sound source. Each model produces a calibration signal, based on the SSL estimate, assuming that the robot is in the environment in which that model trained. In the field, the robot orients its camera toward the sound source using the calibrated estimate, to obtain the ground truth sound source position, from which an error is computed for each model. The likelihood that each model is the best suited to calibrate in the current environment is then computed, and from this a softmax function is used by the *Responsibility Estimator* (RE) to compute a responsibility λ_i for each model

$$\lambda_i = \frac{e^{-|\theta_t - \theta_i|^2/\sigma^2}}{\sum_{j=1}^{n} e^{-|\theta_t - \theta_j|^2/\sigma^2}} \tag{1}$$

where θ_t is the ground truth azimuth, θ_i is the estimate produced by the ith model, n is the number of estimates (models) and σ is a scaling factor which is tuned by hand as in [5]. The calibrated outputs of the models are then combined in proportion to their responsibility values to produce an overall calibration signal. In this way, the system can select a set of models (rather than the best single model) to best calibrate the SSL estimate, including allowing adaption to novel contexts.

3 Proposed system

As mentioned in section 2.4, computation of the responsibility signals, required to apportion the calibration effort of the models, relies on the availability of the ground truth. As already discussed in section 1, the ground truth may be unavailable. Even where it is available, it cannot be determined until after the robot has oriented its camera toward the sound source. MOSAIC includes feed-forward selection of models through prediction of the responsibilities based on contextual signals, rather than sensory feedback. Contextual signals are derived from the environment to form a prior prediction of responsibility:

$$\lambda_i = \frac{\lambda_{pi} e^{-|\theta_t - \theta_i|^2/\sigma^2}}{\sum_{j=1}^{n} \lambda_{pj} e^{-|\theta_t - \theta_j|^2/\sigma^2}} \tag{2}$$

where λ_{pi} is the predicted (prior) responsibility of the ith model. Contextual signals could be of any form, auditory, visual, tactile and so on, and in this study are derived from the audio stream itself, using just one feature, the mean zero crossing rate of the audio signal. This feature was chosen for computational simplicity and alone proved sufficient in the experiments conducted here. However, a range of features could have been used, and more challenging real-world environments may require a different set of features. Zero crossing rate is the ratio of the signal zero crossings to the number of audio samples over the analysis frame, and provides a rough indication of the frequency content of the signal (so will tend to be higher where higher frequencies dominate, especially "noisy" signals). Features were extracted using an adapted version of the *Audio Analysis Library* [18]. The RP is shown in the context of the multiple-models calibration system in figure 2. There is one RP associated with each model in the system, and it learns to predict the responsibility of the model based on contextual signals from the environment. The RP based on the adaptive filter model is shown in figure 3. Features are analysed into parallel fibre signals, based on the value of the feature, such that a low value of the feature would activate a parallel fibre toward one end of the array (with the value of the feature), while a high value would activate a fibre toward the other end of the array. This was chosen as an approach to be close to the use of the cerebellar model that calibrates the

SSL estimate as described in [2], in which the parallel fibres transmit activity on the robot's audio map. The output of the adaptive filter is a prediction of the associated model's responsibility $\hat{\lambda}$ and is the sum of the parallel fibre signals (the feature value) multiplied by the parallel fibre-purkinje cell weights

$$\hat{\lambda}_i = \sum_{i=0}^{n} w_i p_i \tag{3}$$

where p_i is the ith parallel fibre signal and w_i is the corresponding weight. The parallel fibre-purkinje cell weights are updated according to the covariance learning rule [19] as shown in figure 3, with the teaching signal based on the overall responsibility signal, as shown in figure 2.

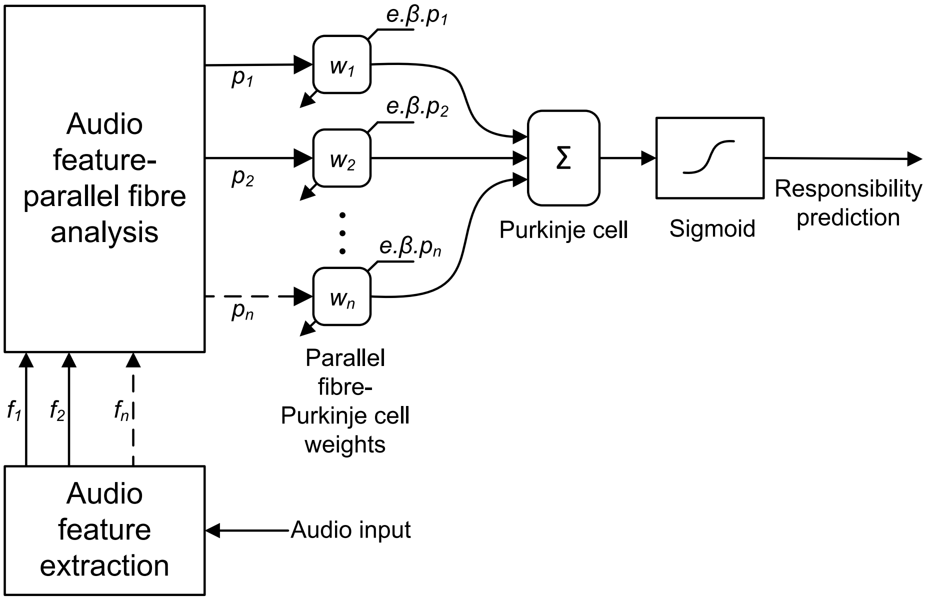

Fig. 3. Cerebellar implementation of the responsibility predictor.

4 Method

Matlab was used to control experiments and for implementation of algorithms. Two microphones (Audio-Technica ATR-3350 omnidirectional condenser lavalier) with an inter-microphone distance of 0.25m were mounted in free field at either end of a horizontal bar, itself mounted on a Pan and Tilt Unit (PTU). The microphones were connected to a computer using a M-Audio MobilePre USB audio capture unit with a sampling rate of 44100Hz. A sound source (Logitech Z150 Speaker) was positioned at a distance of 0.4m from the robot (figure

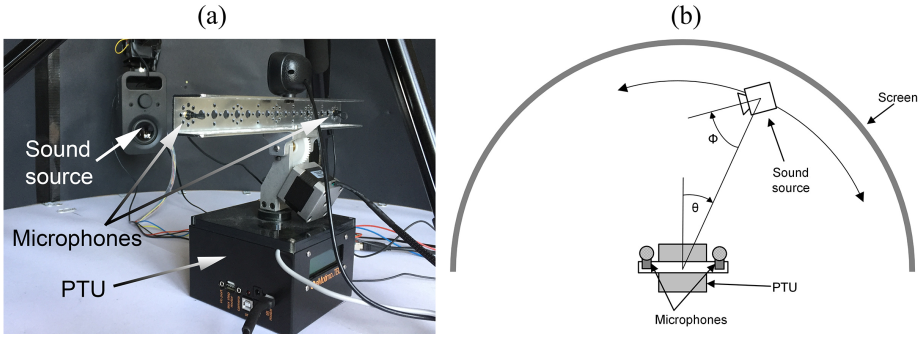

Fig. 4. (a) The experimental arena. (b) Plan view.

4) and was connected to the computer sound card. The sound source could be placed at various azimuths θ (figure 4b), using a tripod arrangement with a geared stepper motor, and could be rotated on its axis (angle ϕ as shown in figure 4b) using a second stepper motor to generate different acoustic contexts. The robot operating in the field would rotate to orient its camera toward the assumed sound source location to determine the ground truth sound source position. For convenience, the ground truth was taken directly from the odometry of the experimental apparatus, and the robot remained stationary. Cerebellar calibration models were trained as in [2] and used to generate target likelihood values using randomly selected samples of the recorded audio data, with corresponding audio features being generated from the same audio samples, in each of three different acoustic contexts.

In MOSAIC, the RP is trained with the posterior responsibility value as a teaching signal. Therefore, at each training iteration the partially trained RP was itself used to make a prediction of the responsibility to be combined with the target likelihood using equation 2 in generating a target posterior responsibility.

Learning rate β, number of parallel fibres, sigmoid shape and number of learning iterations were tuned by hand to obtain a good performance of the RP, which was determined through localization error. Compared to the cerebellar calibration models described in [2], larger values of β and number of training iterations were required to achieve satisfactory performance of the RP. The robot head was presented with a sequence of acoustic contexts each consisting of 5 trials with randomly selected azimuths. In each trial, a 1 second duration Gaussian noise stimulus was used and the SSL unit generated an estimate of azimuth. Each model's calibration output was compared to the ground truth position of the sound source to generate a prediction error for that model. This was carried out in the next trial, because in the field, ground truth would be determined only after the robot had oriented its camera toward the sound source. 10 runs of each experiment were conducted to obtain performance statistics (mean squared error and accuracy rate as percentage of trials in which the absolute error was less than 5^o).

5 Results

Figure 5a shows the responsibility signals of the system as it progressed through trials in contexts in which the calibration models had been pre-trained. By definition, the RPs had to be trained to predict the responsibilities of the calibration models in the same contexts. The blue curves show the normalized likelihood values, that is, the posterior responsibilities, generated after the ground truth becomes available, *without* being combined with RP output, to highlight the behaviour of the RE alone. These posterior responsibilities show the models dominating the responsibility in the context in which they learned (for example, model 1 learned in context 1). It can be observed that there is a delay of one trial before the RE responds to a change in context, as the responsibility values cannot be updated until after the ground truth becomes available in the next trial. Solid orange curves show the outputs of the RPs. The RP output is similar in shape to that of the RE alone, but because the RP is driven by contextual signals derived from the audio stream, it can update its prediction in response to a change in context before the ground truth becomes available. The broken red curve shows the overall responsibility computed using equation 2, so that it is the result of combination of RP output and likelihoods before input to the

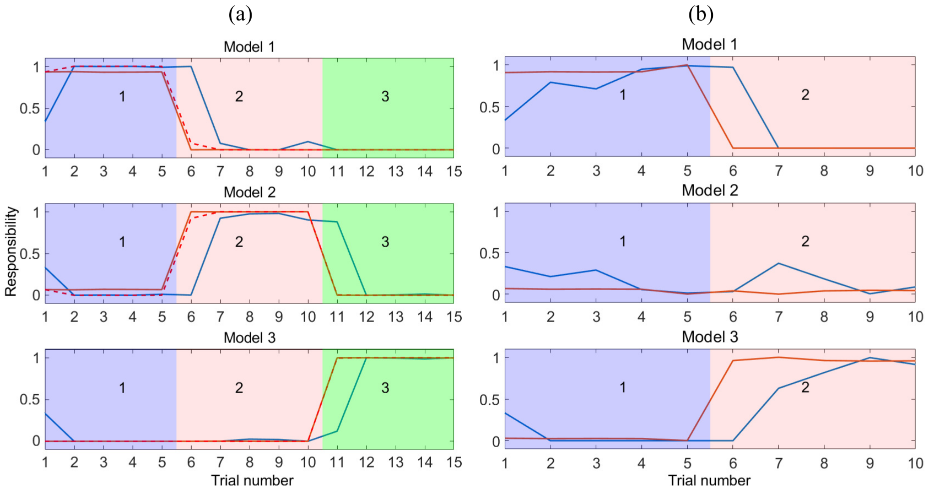

Fig. 5. Responsibility signals as the system progressed through trials. Each trial represents a randomly selected azimuth in one of a number of different contexts, indicated by the coloured boxes. (a) Learned contexts: Context 1 (blue) is ϕ=90° left; context 2 (red) is ϕ=0°; context 3 (green) is ϕ=90° right. The blue curve shows the output of the RE alone, the orange curve shows the output of the RP, and the red broken curve shows the overall responsibility calculated according to Equation 2. (b) Novel contexts: Context 1 (blue) is ϕ=72° left; context 2 (red) is ϕ=72° right.

RE. The RP output closely follows the combined responsibility. Rows 1 and 2 of table 1 show that the performance of the system with the presence of the RP (row 2) is improved over that without (row 1).

The system was tested in contexts that had not been previously experienced. The contexts were chosen to have characteristics that fell intermediate to those in which the calibration models had been trained, as the MOSAIC framework is unable to generalize to contexts that fall outside the learned state space [5]. Figure 5b shows that the system without the RP is able to generalize to the novel contexts, and the the RPs appear able to predict this generalization reasonably well. Rows 3 and 4 of table 1 show that the RP improves the performance in novel contexts.

As mentioned in section 1, computation of the responsibility signals depends on the availability of the ground truth, which may not always be available. The ground truth was made to be unavailable in trial 6 (roughly half-way through the experiment), and remained unavailable throughout the remainder of the experiment. Rows 5 and 6 of table 1 show that the performance with the RP present is improved over that relying on the RE alone. Further, row 7 shows that where the RP alone was used to provide the overall responsibility when the ground truth was unavailable, the performance was comparable with that of the system where the ground truth is available in all trials. However, it should be borne in mind that relying on the RP alone depends on a fully trained RE against which the RP was able to learn before the ground truth became unavailable. Figure 6a shows the profile of responsibilities in this scenario.

The experiment was repeated with a mis-classification of the context by the RPs as discussed in section 2.3. During context 2, the RPs were presented with audio stimulus recorded for context 3 instead of that for context 2, whilst the calibration models themselves were presented with the correct audio recording from context 2. Figure 6b shows that the RPs (orange curve) mis-classify the context, so that, for example, the RP associated with model 3 predicts dominance by that model in context 2 as well as context 3, as would be expected. It can be observed that the RE (blue curve) causes a posterior correction of the overall responsibility, shown by the profile of the red broken curve.

Table 1. Localization performance. N=150. Accuracy is percent less than 5° error

Method	Accuracy	MSE (degrees^2)
1. Combined models without RP	93%	5.5
2. Combined models with RP	100%	1.12
3. Combined models without RP in novel contexts	90%	9.3
4. Combined models with RP in novel contexts	95%	5.7
5. Ground truth missing from trial 6	71%	19.9
6. Ground truth missing with RP	83%	14.5
7. RP alone providing responsibility signals from trial 6	99%	1.7

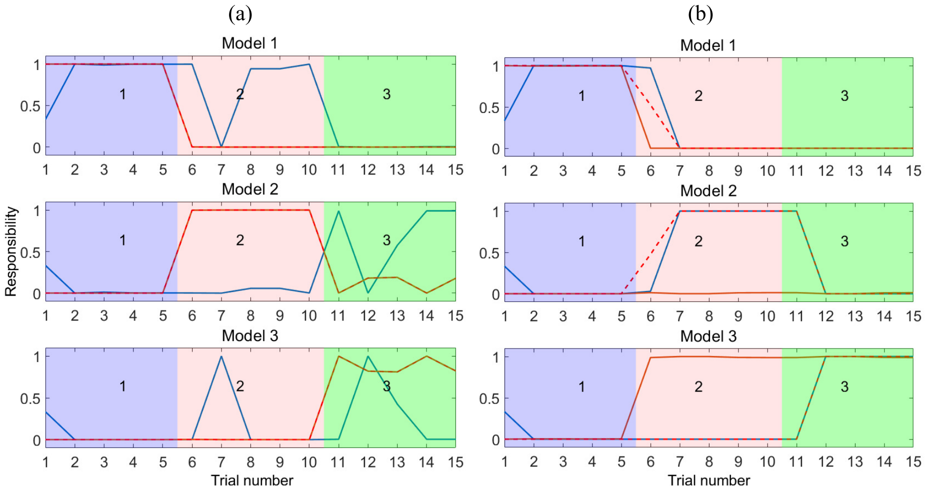

Fig. 6. Responsibility signals as the system progressed through trials. Each trial represents a randomly selected azimuth in one of three different contexts, indicated by the coloured boxes. Context 1 (blue box) is ϕ=90° left; context 2 (red box) is ϕ=0°; context 3 (green box) is ϕ=90° right. The blue curve shows the output of the RE alone, the orange curve shows the output of the RP, and the red broken curve shows the overall responsibility calculated according to Equation 2. (a) Missing ground truth (b) RP mis-classification.

6 Conclusion and discussion

The RP based on the adaptive filter model of the cerebellum was able to successfully predict the responsibility values of the cerebellar models in the acoustic contexts presented, improving the performance compared to that without the RP. It did this through analysis of a single feature extracted from the audio stream allowing the system to predict the responsibilities of the models before the ground truth becomes available. The RP also improves the performance of the system with prolonged absence of the ground truth, allowing the robot to continue to calibrate its SSL even where it moves to a new acoustic environment, which is not addressed in the MOSAIC literature. The system was also able to predict, to a limited extent, the generalization of the models to novel contexts. Finally, it was shown that the RE is able to make a posterior correction to the responsibility values in the event that the RP mis-classifies the context.

References

1. H. G. Okuno, K. Nakadai, "Robot audition: Its rise and perspectives", 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, pp. 5610–5614, 2015.

2. Baxendale, M. D. and Pearson, M. J. and Nibouche, M. and Secco, E. L. and Pipe, A. G., “Audio Localization for Robots Using Parallel Cerebellar Models”, IEEE Robotics and Automation Letters, vol. 3, no. 4, p. 3185–3192, 2018.
3. J. Porrill, P. Dean, and J. V. Stone, “Recurrent cerebellar architecture solves the motor-error problem”, Proceedings of the Royal Society B: Biological Sciences, vol. 271, no. 1541, pp. 789–796, 2004.
4. D. M. Wolpert, R. C. Miall, and M. Kawato, “Internal models in the cerebellum”, Trends in Cognitive Sciences, vol. 2, no. 9, pp. 338–347, 1998.
5. D. M. Wolpert and M. Kawato, “Multiple paired forward and inverse models for motor control”, Neural Networks, vol. 11, no. 78, pp. 1317–1329, 1998.
6. Y. Demiris and B. Khadhouri, “Hierarchical attentive multiple models for execution and recognition of actions”, Robotics and Autonomous Systems, vol. 54, no. 5, pp. 361–369, 2006.
7. M. Johnson and Y. Demiris, “Abstraction in Recognition to Solve the Correspondence Problem for Robot Imitation”, in Towards Autonomous Robotic Systems, TAROS, Essex, UK, pp. 63–70, 2004.
8. Baumann, O., Borra, R., Bower, J., Cullen, K., Habas, C., Ivry, R., Leggio, M., Mattingley, J., Molinari, M., Moulton, E., Paulin, M., Pavlova, M., Schmahmann, J. and Sokolov, A., “Consensus Paper: The Role of the Cerebellum in Perceptual Processes”, Cerebellum, vol. 14, no. 2, p. 197–220, 2015.
9. Snider, R. S. and Stowell, A., “Receiving areas of the tactile, auditory and visual systems in the cerebellum”, Journal of Neurophysiology, vol. 7, p. 331–357, 1944.
10. S. Argentieri, P. Danès and P. Souères, “A survey on sound source localization in robotics: From binaural to array processing methods”, in Computer Speech & Language, vol. 34, no. 1, pp. 87–112, 2015.
11. J. Davila-Chacon, S. Magg, L. Jindong, and S. Wermter, “Neural and statistical processing of spatial cues for sound source localisation”, Neural Networks (IJCNN), The 2013 International Joint Conference on, pp. 1–8, 2013.
12. K. Youssef, S.Argentieri, and J. L. Zarader, “A binaural sound source localization method using auditive cues and vision”, Neural Networks (IJCNN), Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 217–220, 2012.
13. Fujita, M., “Adaptive filter model of the cerebellum”, Biol Cybern., vol. 45, no. 3, pp. 195–206, 1982.
14. P. Dean, J. Porrill, C. F. Ekerot, and H. Jorntell, “The cerebellar microcircuit as an adaptive filter: experimental and computational evidence (report)”, Nature Reviews Neuroscience, vol. 11, no. 1, p. 30, 2010.
15. H. Imamizu and M. Kawato, “Brain mechanisms for predictive control by switching internal models: implications for higher-order cognitive functions”, Psychological research, vol. 73, no. 4, pp. 527–544, 2009.
16. M. Haruno, D. M. Wolpert and M. Kawato, “MOSAIC Model for Sensorimotor Learning and Control”, Neural Computation, vol. 13, no. 10, pp. 2201–2220, 2001.
17. T. Assaf, E. D. Wilson, S. Anderson, P. Dean, J. Porrill, and M. J. Pearson, “Visual-tactile sensory map calibration of a biomimetic whiskered robot”, in 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2016, Conference Proceedings, pp. 967–972.
18. T. Giannakopoulos and A. Pikrakis, “Introduction to audio analysis: a MATLAB approach”, First ed., Amsterdam: Academic Press, 2014.
19. T. J. Sejnowski, “Storing covariance with nonlinearly interacting neurons”, J Math Biol, vol. 4, no. 4, pp. 303–21, 1977.