CALIBRATION OF SOUND SOURCE LOCALISATION FOR ROBOTS USING MULTIPLE ADAPTIVE FILTER MODELS OF THE CEREBELLUM

Mark David Baxendale

A thesis submitted in partial fulfilment of the requirements of the University of the West of England, Bristol, for the degree of Doctor of Philosophy

Bristol Robotics Laboratory, Faculty of Environment and Technology, University of the West of England, Bristol, United Kingdom

February 2020

Abstract

The aim of this research was to investigate the calibration of *Sound Source Localisation* (SSL) for robots using the adaptive filter model of the cerebellum and how this could be automatically adapted for multiple acoustic environments. The role of the cerebellum has mainly been identified in the context of motor control, and only in recent years has it been recognised that it has a wider role to play in the senses and cognition. The adaptive filter model of the cerebellum has been successfully applied to a number of robotics applications but so far none involving auditory sense. Multiple models frameworks such as *MOdular Selection And Identification for Control* (MOSAIC) have also been developed in the context of motor control, and this has been the inspiration for adaptation of audio calibration in multiple acoustic environments; again, application of this approach in the area of auditory sense is completely new. The thesis showed that it was possible to calibrate the output of an SSL algorithm using the adaptive filter model of the cerebellum, improving the performance compared to the uncalibrated SSL. Using an adaptation of the MOSAIC framework, and specifically using *responsibility estimation*, a system was developed that was able to select an appropriate set of cerebellar calibration models and to combine their outputs in proportion to how well each was able to calibrate, to improve the SSL estimate in multiple acoustic contexts, including novel contexts. The thesis also developed a *responsibility predictor*, also part of the MOSAIC framework, and this improved the robustness of the system to abrupt changes in context which could otherwise have resulted in a large performance error. Responsibility prediction also improved robustness to missing ground truth, which could occur in challenging environments where sensory feedback of ground truth may become impaired, which has not been addressed in the MOSAIC literature, adding to the novelty of the thesis. The utility of the so-called *cerebellar chip* has been further demonstrated through the development of a responsibility predictor that is based on the adaptive filter model of the cerebellum, rather than the more conventional function fitting neural network used in the literature. Lastly, it was demonstrated that the multiple cerebellar calibration architecture is capable of limited self-organising from a de-novo state, with a predetermined number of models. It was also demonstrated that the responsibility predictor could learn against its model after self-organisation, and to a limited extent, during self-organisation. The thesis addresses an important question of how a robot could improve its ability to listen in multiple, challenging acoustic environments, and recommends future work to develop this ability.

Acknowledgements

I would like to thank my Director of Studies, Tony Pipe, along with my other supervisors, Mokhtar Nibouche, Martin Pearson and Emanuele Secco, who were a first class supervision team. Doing a doctorate part time has been very challenging, especially as I live 180 miles from the lab, and their support has been second to none.

I wish to thank Ahmad Sheikh, a graduate of UWE, for his contribution to developing the moving sound source apparatus described in Chapter 6, as part of an internship project.

I would also like to thank my line manager, Atulya Nagar, who has supported me throughout the project, providing me generously with time to do the research as well as crucial resources.

I thank my family for their support and encouragement- my wife Sue for her patience and support as I spent any spare moment either away from home doing experiments or sat at the computer; my mother Peggy and my late father Leo, my siblings Martin, Carol, Stephen and Heather as well as my daughter Ellie and son Jacob.

Publications from the thesis

M. D. Baxendale, M. J. Pearson, M. Nibouche, E. L. Secco, and A. G. Pipe, "Self-adaptive Context Aware Audio Localisation for Robots Using Parallel Cerebellar Models," in Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Guildford, UK, July 19–21, 2017, Proceedings, Y. Gao, S. Fallah, Y. Jin, and C. Lekakou, Eds., ed Cham: Springer International Publishing, pp. 66-78, 2017.

M. D. Baxendale, M. J. Pearson, M. Nibouche, E. L. Secco, and A. G. Pipe, "Audio Localisation for Robots Using Parallel Cerebellar Models," IEEE Robotics and Automation Letters, vol. 3, pp. 3185-3192, 2018.

Baxendale M.D., Nibouche M., Secco E.L., Pipe A.G., Pearson M.J. (2019), "Feed-Forward Selection of Cerebellar Models for Calibration of Robot Sound Source Localization," in: Martinez-Hernandez U. et al. (eds) Biomimetic and Biohybrid Systems. Living Machines 2019. Lecture Notes in Computer Science, vol 11556. Springer, Cham

Section 1.7 outlines which parts of the thesis have been published.

Contents

List of figures

List of tables

Nomenclature

α	Sound source angle of elevation
α^H	Hidden-Markov forward probability
β	Cerebellar learning rate
β^H	Hidden-Markov backward probability
γ	Responsibility of HMM-MOSAIC module
ξ	The set of all possible sequences in HMM-MOSAIC
ϕ	Sound source angle of rotation about its vertical axis
θ	Azimuthal sound source position
λ	Responsibility
σ	Softmax scaling factor
ASC	Acoustic Scene Classification
AAL	Audio Analysis Library
AEG	Auditory Epipolar Geometry
CN	Cochlear Nucleus
CASA	Computational Auditory Scene Analysis
DFT	Discrete Fourier Transform
DCN	Deep Cerebellar Nuclei, Dorsal Cochlear Nucleus (depending on context)
ESR	Environmental Sound Recognition
EM	Expectation Maximisation
FPGA	Field Programmable Gate Array
GCC-PHAT	Generalized Cross Correlation with Phase Transform
HRTF	Head Related Transfer Function
HMM	Hidden-Markov Model
IC	Inferior Colliculus
IHC	Inner Hair Cells
ILD	Inter-aural Level Difference (of sound)
ITD	Inter-aural Time Difference (of arrival of sound)
I/O	Input/Output
LMS	Least Mean Square
LSO	Lateral Superior Olive
MSE	Mean Squared Error
MSO	Medial Superior Olive
MFCC	Mel-Frequency Cepstrum Coefficient
MOSAIC	MOdular Selection And Identification for Control
NN	Neural Network
OHC	Outer Hair Cells
PTU	Pan-and-Tilt Unit
RE	Responsibility Estimator
RP	Responsibility Predictor
SFS	Sequential Forward Selection
SNN	Spiking Neural Network
SOM	Self-Organising Map
ST	Scattering Theory
SSL	Sound Source Localisation
SC	Superior Colliculus
SoC	System on Chip
SOC	Superior Olivary Complex
USB	Universal Serial Bus
VCN	Ventral Cochlear Nucleus
VLSI	Very Large Scale Integration

Chapter 1 Introduction

1.1 Background

The main motivation for this thesis was to develop a system that could support a robot in locating sources of sound in its environment, in particular a rescue robot that is attempting to locate survivors in the aftermath of a disaster. Typically, vision is used as the primary sense in the field because it is generally considered to be the most reliable and robust sense available to a robot [1]. However, in extreme or challenging environments such as a disaster site, vision could be impaired by, for example, particles in the air, or the vision sensor(s) could even become impaired through damage. In these situations, other senses, such as audio, could take over, or at least play a more prominent role, until the more reliable sense once again becomes available.

The thesis focuses on *Sound Source Localisation* (SSL), and in particular, determining the azimuth direction of arrival of sound. The distance to the sound source and its elevation (which, together with the direction of arrival define the location of a sound source in 3 dimensional space; see Figure 4) are not considered. However, as discussed below and later in this thesis, the proposed system should lend itself to including more extensive schemes such as those estimating distance and elevation.

SSL techniques are well established [2] with some quite sophisticated and robust techniques now available. The thesis adopts a basic approach using *Interaural Time Difference* (ITD) of arrival of sound, which is described in Section 2.3. There are a number of different approaches to SSL and the thesis specifically adopts *binaural* SSL for reasons outlined in Section 2.3. Such a basic approach is sufficient to demonstrate the utility of the proposed system introduced in the thesis, and the design of the system is such that a more sophisticated SSL algorithm could be easily substituted for the one used.

A problem with established SSL techniques is that they are generally designed to operate in well-defined and controlled environments. For a robot operating in the field, especially in disaster situations such as inside a collapsed building, errors will be introduced into the SSL process, due to complex environmental acoustics that are not easily determined, and this error may depend on the azimuth sound source position, perhaps in a non-linear fashion. Even where an SSL system is designed to cope with such errors, they are generally only designed to do so in a particular environment. If the robot moves to a

different environment, it is likely that the SSL system designed to compensate for errors introduced by the first environment will be unable to compensate for the errors introduced by the new environment which could display quite different acoustic properties.

There has been growing acceptance that the brain makes use of internal models for motor control and that they are likely to be located in the cerebellar cortex [2, 3], and also that they play a role in non-motor functions, including in perceptual processes [4, 5]. Despite this, there has been no investigation, to the knowledge of the supervision team, into the use of cerebellar models in robot audition. The thesis develops a model of the cerebellum that adapts, or calibrates, the output of an SSL system to compensate for errors introduced by the acoustic properties of the environment that the robot is operating in. A cerebellar model has to learn, through repeated operation in the given environment, with a teaching signal based on its performance error, to calibrate the output of the SSL system. The adaptive filter model of the cerebellum is a type of NN (although later, in Chapter 8, a distinction is made between "standard" NNs and the cerebellar models developed in this thesis). Like a NN it learns through the updating of synaptic weights. It learns to calibrate the SSL system through repeated stimulation by sound from random azimuths, estimating the direction of arrival of the sound, comparing this to the true direction, then updating the synaptic weights (in this case, parallel fibre-Purkinje cell weights as described in Chapter 3) in such a way as to reduce the error in estimation.

The project specifically builds on previous work at Bristol Robotics Laboratory (BRL) and the University of Sheffield; the *Bioinspired Control of Electro-Active Polymers for Next Generation Soft Robots* (Bella) project [3]. That project used an adaptive filter model of the cerebellum to calibrate a somatosensory (whisker) map on a robot using visual input [4, 5]. The basic thrust of the current project is to add audio input to such a system, and to extend the approach to multiple cerebellar models. The project however does not use the Bella robot (known as Bellabot), although software from that project has been adapted for use in the current project; rather, a version was developed which replaces the whisker sensor system with audio. This is explained in more detail in Chapter 4.

Regarding the multiple-models approach to calibrating the SSL for multiple acoustic environments, there is a problem of how to select the appropriate set of models in a particular acoustic environment. There are a number of approaches, mainly developed in the context of motor control as mentioned above, and the one this thesis focuses on is

MOdular Selection And Identification for Control (MOSAIC) [6]. MOSAIC is based on the existence of multiple models, internal to the brain of an animal (or processing unit of a robot) that represent the external world, allowing the prediction of the consequences of some action by the animal or robot. The MOSAIC system refers to modules rather than models, with each module containing separate, specialised models for prediction and control. The system uses the prediction errors, determined through sensory feedback, to select the best modules for future control. This can be applied to the SSL calibration system, so that the best calibration model, or set of models can be selected in a given acoustic environment, based on how well each is able to calibrate the SSL estimate in that environment. The calibration effort of the models can then be combined in proportion to how well each model is able to calibrate the SSL output. A model with a much smaller prediction error (and hence which appears to be the "best" model by a good margin) will tend to dominate control (or rather calibration, in the context of this thesis), with some contribution from the next best model and so on. An environment that elicits a less distinctive response from the models will see control/calibration more equally shared between models. Hence, the focus of the thesis is on developing a system which is able to adapt *between environments*, rather than show a particular performance in a single environment. The system takes a SSL estimate as one of its inputs (in the case of responsibility prediction, covered in Chapter 8, contextual signals form a second input) and as such the performance of the system will depend on the reliability of the underlying SSL system. SSL systems are particularly vulnerable to environmental acoustics, especially noise and reverberation, and the error in the SSL estimate could be so severe that even calibration as proposed in this work may break down. The decoupling of the system from the underlying SSL system however, means that a different, perhaps more robust system can be "plugged in" as such systems develop.

An early motivation for this work was audio-visual integration, along with a biomimetic modelling of the auditory periphery. The focus moved away from this and towards a multiple-models inspired approach to SSL calibration. Sensor fusion is still in the background, even though it is not used in this thesis; vision is (potentially) used as a means of determining the ground truth sound source position. Full modelling of the auditory periphery was not considered necessary, with a simple SSL algorithm being sufficient to demonstrate the overall idea. A more full modelling of the auditory periphery

could of course be re-introduced, particularly in the context of substituting more sophisticated SSL, and as mentioned in Section 2.2.1 could be included in future work.

1.2 Novelty of the thesis

Two key concepts are presented in the thesis as the basis for novelty. The first is calibration of SSL using an adaptive filter model of the cerebellum, and the second is the development of a multiple-models inspired approach ("borrowed" from the field of motor control) to selecting an appropriate set of adaptive filter models of the cerebellum, each of which has learned to calibrate the SSL system in a particular acoustic environment, or context, to suit the context that the robot finds itself in. A further, secondary area of novelty that emerged later in the project is the development of a *Responsibility Predictor* (RP) that is also based on the adaptive filter model of the cerebellum, rather than the more conventional function fitting Neural Network (NN, Section 8.5.5) found in the literature.

1.3 Aims

The main aim of the thesis is to develop a system capable of calibrating an SSL system for multiple acoustic environments using multiple adaptive filter models of the cerebellum.

A secondary aim of the thesis is to further demonstrate the utility of the so-called "cerebellar chip" [7-9] in a new area.

1.4 Objectives

1. Development of a binaural SSL algorithm
2. Development of cerebellar calibration of SSL
3. Development of a multiple-models approach to cerebellar calibration
    a. Development of a context estimation sub-system
    b. Development of a multiple model based SSL calibration sub-system
    c. Development of a responsibility prediction sub-system
    d. Investigation of de-novo learning of the cerebellar models
4. Design and performance of SSL experiments based on each developed system or sub-system.

1.5 Research questions

1. Is it possible to apply cerebellar calibration to SSL?

2. Is it possible to implement a multiple-models inspired architecture that will select an appropriate cerebellar calibration model, or set of models, in different acoustic contexts?
3. Can a multiple-models inspired audio calibration system self-organise?

1.6 Proposed system and structure of the thesis

The thesis draws on three key areas, robot audition, cerebellar models and multiple models. Chapter 2 introduces robot audition with an emphasis on SSL; Chapter 3 covers the cerebellum with an emphasis on the adaptive filter model of the cerebellum; Chapter 5 is a treatment of multiple models with an emphasis on the MOSAIC scheme. Chapter 4, Chapter 6, Chapter 7 and Chapter 8 describe the work carried out to develop the proposed system, which is depicted in Figure 1. In this figure, the main elements are emphasised using blue tint boxes which also indicate the chapters that cover those elements. SSL calibration (Chapter 4) using the adaptive filter model of the Cerebellum addresses research question 1 and the model developed is an adaptation of that used in a previous study to calibrate whisker input to a robot platform [4]. This aspect of the system learns to compensate for errors in the SSL estimate in particular acoustic environments. Figure 1 shows multiple models for completeness, although Chapter 4 itself is based on a single model, as was presented in [4]. Responsibility estimation is covered in Chapter 6 and Chapter 7, and is a fundamental aspect of the development of the work of Chapter 4 into a multiple-models inspired system. Chapter 6 focuses on the use of the responsibility estimator to identify the acoustic context while Chapter 7, the key chapter of the thesis, combines the multiple model outputs to improve the SSL calibration in multiple acoustic environments. Chapter 8 does not strictly address the research questions and was added as the author wished to investigate responsibility prediction, mainly for completeness, although this turned out to produce a quite important finding in terms of missing ground truth. The same can be said for Chapter 9, de-novo learning of the models, which, although it addresses one of the research questions, was included for completeness. A weak area of the thesis is how the system could be moved towards a real world environment, and Chapter 10 starts to investigate this as a possibility, and issues identified are discussed in Chapter 11 where future work is identified.

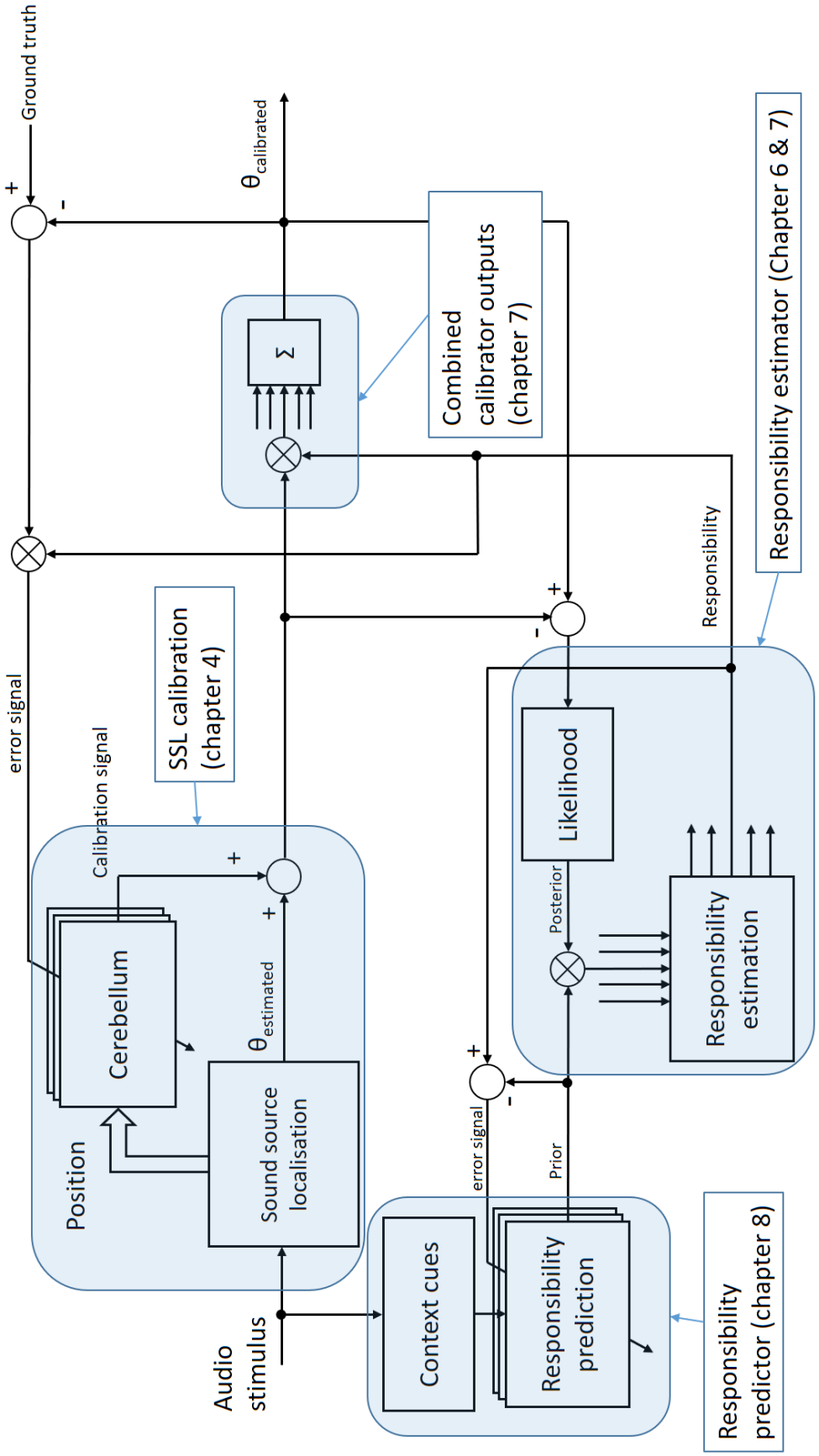


Figure 1. The overall architecture of the proposed system. The blue boxes indicate key components to which a chapter is wholly or partly dedicated.

1.7 Publications from the thesis

The work in Chapter 6 was published in *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017*.

The work in Chapter 7 and part of the work in Chapter 8 and part of the work in Chapter 10 was published in *IEEE Robotics and Automation Letters*.

A further part of Chapter 8 was published in *Biomimetic and Biohybrid Systems: 8th International Conference, Living Machines 2019*.

Chapter 2 The auditory system and robot audition

2.1 Introduction

There is a need for autonomous mobile robots to use a variety of senses to navigate or locate entities in unstructured environments. Typically, vision is used to locate objects in the robot's environment, however, this can break down where vision is obscured. This is a particular issue in disaster situations such as those described in Section 1.1. A number of attempts have been made to allow a robot to navigate by sound alone [2], however these systems are typically set up in a specific acoustic environment and can break down when the robot moves to a new environment.

2.2 The biological auditory periphery

The structure of the animal auditory periphery will be referred to at various points throughout the thesis, and, although only simplified aspects of the auditory system are included, in particular, those concerned with SSL, a brief treatment is useful to set part of the context in which the thesis is positioned. The system described here is mainly based on the mammalian biology, as both the avian and reptilian auditory pathways possess some significant differences. Treatment of the mammalian auditory pathway however, brings out some basic concepts that are relevant to *Robot Audition*, discussed in Section 2.3.

The primary organ of the inner ear is the *cochlea* which appears early in the auditory pathway. The basilar membrane that runs the length of the cochlea holds the auditory nerve, which consists of rows of hair cells which transduce the basilar membrane vibration (itself caused by vibration of the *tympanic membrane,* that is, the “eardrum”, and the *ossicles* due to variations in air pressure caused by sound) into an electrical waveform. There are two distinct types of hair cell known (because of their orientation) as *Inner Hair Cells* (IHC) and *Outer Hair Cells* (OHC). The former plays the major part in the conversion process [10]. A low or high frequency stimulus will cause a peak in vibration amplitude toward the apex and base of the basilar membrane respectively. This in turn will give rise to a peak nerve fibre response to each component at a place along the basilar membrane that is frequency-specific. The function of the cochlea seems to be to analyse the audio stimulus into a number of frequency bands (many thousand in the mammalian cochlea), a function often known as the *auditory filter*.

A somewhat simplified schematic of the auditory pathway is shown in Figure 2. The auditory nerve connects to the *Cochlear Nucleus* (CN). The CN includes the *Dorsal Cochlear Nucleus* (DCN[1]), which appears to be linked to *Inter-aural Level Difference* (ILD, Section 2.3) processing and the *Ventral Cochlear Nucleus* (VCN) which appears to be linked to *Inter-aural Time Difference* (ITD, Section 2.3) processing. The outputs of the CN connect to the *Inferior Colliculus* (IC), with some outputs connecting via the *Superior Olivary Complex* (SOC), which plays a key role in sound localisation, and itself includes the *Medial Superior Olive* (MSO), which is linked with ITD processing to determine azimuth, and the *Lateral Superior Olive* (LSO) which is associated with ILD processing to also determine azimuth. There are connections from the IC to the *Superior Colliculus* (SC), which plays a key role in the integration of the senses, but with most connections going to the auditory cortex via the *Auditory Thalamus*.

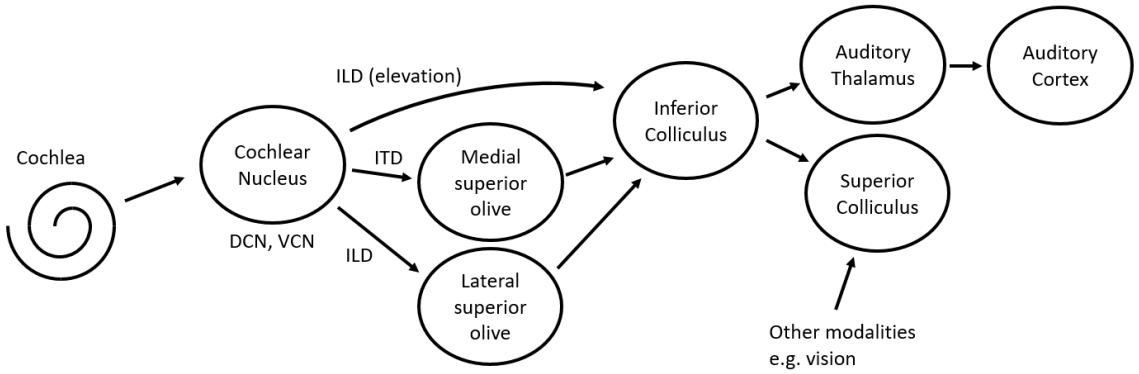


Figure 2. The auditory pathway. DCN[1] is Dorsal Cochlear Nucleus, VCN is Ventral Cochlear Nucleus.

2.2.1 Modelling the auditory periphery

Although the auditory periphery was not modelled as part of the thesis (except on a very limited scale to implement an SSL algorithm), it is quite possible that this could be included in future work, and it is dealt with here briefly, for completeness. The auditory filter in particular has been extensively modelled, mainly for speech processing, but also more recently for Robot Audition, notably in the *Auditory Toolbox* [11] and the *Auditory Modelling Toolbox* [12], both of which provide Matlab libraries (and Octave in the case of [12]) for the simulation of this and other functions of the auditory pathway.

[1] DCN is used elsewhere in the thesis with different meaning

The place coding of the basilar membrane can be modelled electronically as a bank of filters. Models fall predominantly into two categories: a series of electronic filters proposed by Lyon [13] and a parallel bank of filters [14], the latter based on gammatone filters, so called because of the filter impulse response that resembles a gamma-modulated tone, and is widely accepted to reflect the response of the auditory filter; Figure 3 shows the simulated impulse response of a gammatone filter in Simulink. Filters toward the high frequency end of the bank are characterised by higher frequency tones and shorter decay times versus lower frequency tones and longer decay times toward the low frequency end.

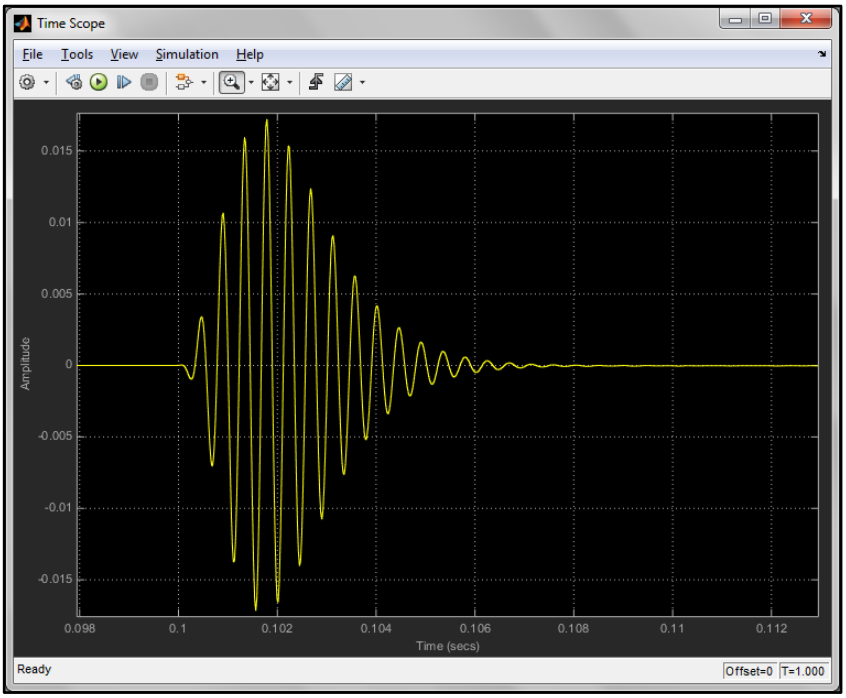


Figure 3. Impulse response of gammatone biquad filter in Simulink.

Approaches to modelling the auditory periphery fall into analogue or digital implementations. Digital implementations are of course commonplace [15-17]. Lyon's own implementation of his model was analogue [18] and analogue implementations continue to enjoy popularity [19-21], especially with the development of *neuromorphic engineering*, proposed by Mead in 1989 [22] in which *Very Large Scale Integration* (VLSI) circuits mimic cells in the central nervous system. Initially these were based on analogue devices but neuromorphic systems now also include digital technologies. An example of this is the *SpiNNaker Project*, a large scale, parallel, spiking NN computing

platform [23]. There have been a number of examples of neuromorphic auditory modelling [24-28].

2.3 Robot Audition

Robot Audition is a relatively recent area of research (around 18 years old at the time of writing) focusing on robot hearing [29], and has developed from, and is related to other areas such as *Computational Auditory Scene Analysis* (CASA) [30], and speech recognition. The newness of this area of research is indicative of the fact that robot vision has dominated research in robot senses, which is understandable given the large amount of data available from devices (image capture devices such as cameras) which are generally low cost and straightforward to use (but which can require significant computing resources to process). Also, the format of the information available through vision sensors makes tasks such as navigation and object localisation relatively straightforward compared to, for example, performing the same task using sound. This is because the sensing array associated with a vision sensor such as a camera has elements that map directly on to positions in the visual scene. This is not possible with audio, especially with the biological auditory system which uses two sensors (ears) and codes sounds from the early stages of the auditory pathway on the basis of frequency content. However, the growing need for robots to operate in the field, that is, in unstructured and unpredictable environments makes the reliance on vision alone sometimes infeasible, especially in extreme environments such as in a disaster situation, and taking into account the processing overhead associated with visual input on a mobile robot which may have limited computing resources.

Robot Audition consists of three key functions: SSL, separation of the sound sources in the audio scene and source recognition [31]. SSL is a low level function forming the basis for higher level functions of source extraction, source recognition and lastly scene reconstruction which is a high level description of the auditory scene [32]. So, although the quite narrow focus of this thesis is on locating sound sources in the robot's environment, as an end in itself, successful calibration of this task could have significant impact on dependent, higher level functions of Robot Audition.

These functions underpin a number of tasks that a robot might be required to undertake such as speaker and emotion recognition. Much of this research (and hence the tasks

focused on) have been motivated by the desire to improve human-robot interaction, and in particular to make this seem more natural.

Robot navigation by sound alone has been well researched [33-36], however the systems described have typically been tested in a specific acoustic environment. Although there are examples of SSL systems that can adapt to different acoustic environments [19, 37-42] it is not clear how well they would perform in a new environment that has not been previously experienced by the robot.

Figure 4 shows the 3D audio map of the robot defined by azimuth θ, elevation α and distance to source. SSL is the process of estimating these parameters that will locate the source of a sound in 3D space, with respect to the robot head, which is located at the centre of a sphere. The thesis considers only 2D (azimuthal) SSL. For convenience, it also only considers sounds that emanate from in front of the robot, rather than behind. Front/back confusion is a particular problem in SSL, and is considered here to be a problem that might be resolved by the appropriate design of the underlying SSL system rather than the system that is the subject of this thesis. As the proposed system includes an adaptation of a previously developed system that calibrated a 2D whisker map, it ought to be possible to extend the SSL calibration system to 2 dimensions, e.g. azimuth and elevation.

The SSL field has been very active over the last 20 years and SSL systems have become quite robust and sophisticated. However, this is largely at the cost of simplicity and size, usually with systems utilising an array of multiple microphones. The thesis specifically eschews microphone arrays and focuses on binaural robot audition. As mentioned in Section 1.1 the system developed here is intended to sit on top of an underlying SSL system. The motivation therefore is towards simplicity, low cost and low computational load, so that a basic binaural system is sought. The SSL approach used in this thesis is based on passive binaural localisation using *Inter-aural Time Difference* (ITD) of arrival of sounds [43], with microphones mounted in free field, corresponding to *Auditory Epipolar Geometry* (AEG) [2]. The head related transfer function of the robot head is ignored. If a modular approach is taken, so that the SSL system is not integrated then it should be a simple matter to "plug in" a more sophisticated system (which could even include those utilising microphone arrays, rather than being binaural). The primary auditory cues used in the passive, binaural localisation of sound sources are ITD and

Inter-aural Level Difference (ILD) [43]. ILD relies on acoustic shadowing caused by the head of the animal (or robot); as such it is frequency dependent, and is effective for higher frequencies (greater than around 1500 Hz). On the other hand, ITD cues are limited to lower frequencies due to phase ambiguity as the period of the sound wave becomes comparable to the maximum ITD available for a given sensor or ear separation [43]. Sound from a source to either side of the median plane will reach one or other sensor or ear at different times (e.g. a sound originating from a source to the right of the median plane will reach the right ear or sensor before the left). The ITD has a maximum value of around 660μs at an azimuth of 90° in humans [43], representing an inter-aural distance of around 15cm. This is subject to uncertainty due to environmental influences such as obstruction of the sound source, the acoustic properties of surfaces or damage to or displacement of audio sensors. There are a number of attempts of SSL [20, 26, 44-52], mostly using ITD. Modelling of the ITD is most commonly carried out using the Jeffress model [53].

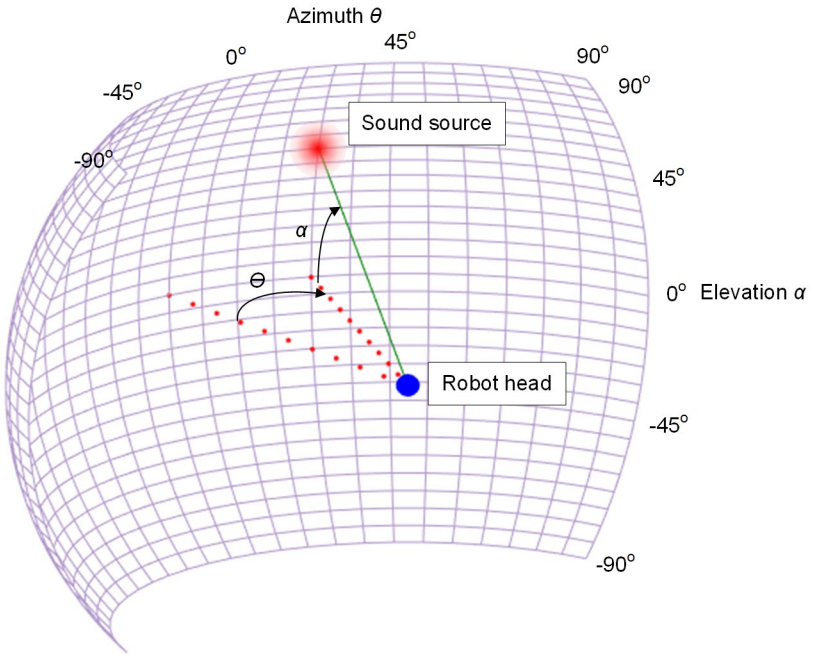


Figure 4. Audio map of sound source location in head-centric space. Reprinted from [54]. © 2018 IEEE.

SSL systems have typically used arrays of multiple microphones [35, 36, 55], however this thesis adopts a binaural approach with a view to the limited resources available on a mobile platform including size, computation and power constraints point to lower

numbers of microphones [32], with two microphones being the minimum number to allow the one-dimensional location of a sound source (i.e. the direction of arrival or azimuth of the sound source in one plane). Those adopting a binaural approach have mostly been under controlled, limited conditions [25, 56-61]. Among those approaches that do use a binaural approach, many use active, rather than passive SSL, where the robot has to move to improve SLL robustness [29, 31, 32].

Other approaches could have been adopted such as Scattering Theory (ST) [25, 56-59], where the robot head is modelled as a sphere. By focusing on AEG the thesis has side-stepped the need to consider the *Head Related Transfer Function* (HRTF). The HRTF is the characteristic response of the acoustic environment that intervenes between the sound source and the inner ear (or microphone) [10]. It will include the acoustic properties of the animal's or robot's head and pinnae. Including the HRTF would involve considerable effort not directly relevant to the central research questions of this thesis. However, the HRTF is still problematic, especially as it needs to be taken into account in the robot's context (or an extensive HRTF database needs to be used) with concomitant inflexibility; the potential for cerebellar calibration to compensate for variation in HRTF among other things may be a rich source of future work.

ITD is the difference in time of a sound reaching each of a binaural sensor array (such as a pair of microphones or an animal's ears). If the sound is at zero azimuth, then as shown in Figure 4, the sound source is directly ahead of the robot and the sound will reach both sensors at the same time, resulting in a value of zero for the ITD. If the sound source changes position to the right of zero azimuth sound from that source will reach the right sensor before the left, as it will have a lower distance to travel o that sensor, resulting in a non-zero ITD. The greater the displacement of the source from zero azimuth, the greater the value of the ITD.

A cross-correlation algorithm has been developed in Matlab to determine the ITD from two microphones and provide an estimate of the azimuth of the location of a sound source. The algorithm uses the Matlab `xcorr` function, which returns the cross correlation between the right and left channels as a function of time difference between the channels, shifting the timebase of one channel with respect to the other by varying amounts. The cross correlation is given by

$$r_{ly} = \sum_{k=0}^{n} R(k)L(k-\tau) \quad (1)$$

where R is the right- and L the left channel audio signal, k is the sample number, n is the current sample and τ is the time lag between left and right channel (the cross correlation is actually found recursively so that the calculation uses the current audio samples and the previous value for r, otherwise it would need to be calculated for all samples at each time step which would quickly become unmanageable). The algorithm finds that time difference which results in maximum similarity between the two channels (maximum correlation value), which corresponds to the time difference of arrival. This is then transformed into an estimated azimuth (Figure 5) by

$$\theta = \frac{180}{\pi} \sin^{-1}\left(\frac{c\tau}{Df_s}\right) \quad (2)$$

where c is the velocity of sound (ms^{-1}), τ is the estimated ITD (s), D is the inter-aural distance (m) and f_s is the audio sampling frequency (Hz). By convention, azimuth values to the right of zero are treated as positive and those to the left are treated as negative.

Results are limited by the resolution of the SSL unit, which varies from +/-1.7º at zero azimuth to +/-5º at +/-70º azimuth. The resolution is affected by the sampling frequency (44100 Hz in this thesis) and inter-microphone distance (0.16m and 0.25m in this thesis-figures above for resolution are for the .25m inter-microphone distance).

A problem with single direction of arrival estimation techniques is that they can break down in noisy or reverberant environments. One approach that is commonly used and which has some robustness to reverberation is GCC-PHAT. Knapp and Carter describe the *Generalized Cross Correlation* (GCC) method [62], which is a cross correlation between the two signals carried out in the frequency domain, with a weighting function.

Generalized Cross Correlation with Phase Transform (GCC-PHAT) uses a weighting function that is the inverse of the cross-correlation which normalizes the magnitudes of peaks in the cross-correlation. This tends to separate out the peaks due to multiple sound sources, providing some robustness in noisy and reverberant environments, making it a commonly used technique. However, GCC-PHAT was not used in this thesis, apart for the purposes of comparison in Chapter 7 and to check that the proposed system could be used with a different SSL algorithm to that used in the thesis.

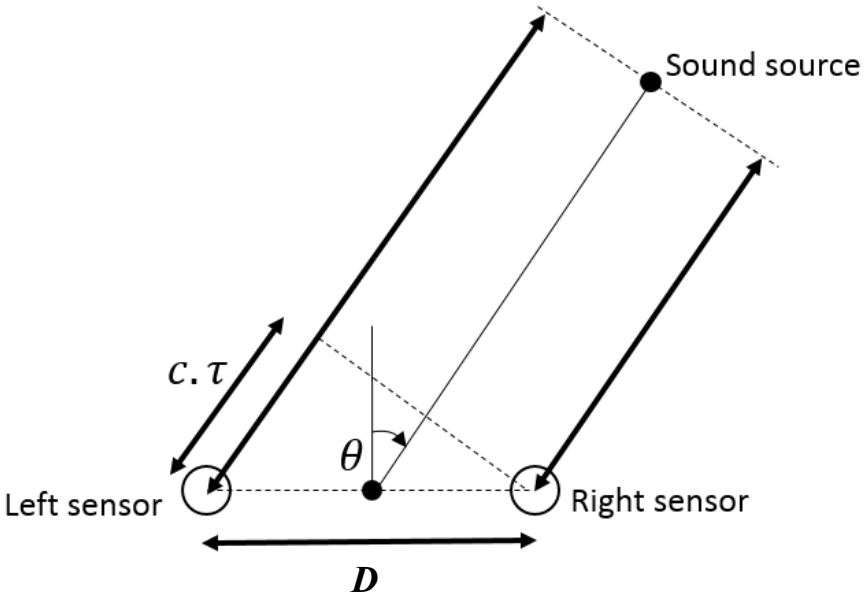


Figure 5. Determining azimuth θ from Inter-aural Time Difference of arrival. c is the velocity of sound, τ is the ITD and D is the inter-aural distance. A planar sound wave is assumed.

2.4 Adaptive SSL systems

A key claim of this thesis is that the proposed system can calibrate an underlying SSL algorithm in multiple environments and select a set of models and blend their outputs to improve the SSL estimate as the robot moves between environments. There are a number of systems that can learn the audio map of a robot, however, there is very little in the literature about systems that will select these learned models from environment to environment [41, 58, 63-65]. There are robot audition systems that use *Acoustic Scene Classification* (ASC- see Section 8.2), sometimes also referred to *Environmental Sound Recognition* (ESR) [66], however these focus on the classification problem itself, and although it has been stated that a robot could switch its mode of operation based on recognition of its acoustic environment [67-70], it is not clear from the literature how ASC or ESR systems are actually used to adjust robot audition functions such as SSL in multiple acoustic environments, as the robot moves between the environments. Indeed, there is no reference whatsoever in the literature to the *blending* of the adjustment or calibration of robot audition functions based on the recognised environment as proposed in this thesis. The thesis draws on work in ASC in the development of an RP.

Chapter 3 The cerebellum

3.1 Introduction

The cerebellum is a densely populated part of the hindbrain of vertebrates and appears to take part in a wide range of functions ranging from fine-tuning of motor control to providing a subconscious sense of agency and self [9]. It has a highly regular structure that turns out to be very versatile in terms of function.

A number of models of the cerebellum have been proposed and that used in this thesis is based on the adaptive filter model of the cerebellum [71, 72], which has proven to be a versatile and robust algorithm in a variety of robotics applications [4, 7, 8].

3.2 Structure of the cerebellum

The cerebellum consists of two main regions, the *Cerebellar Cortex* and the *Deep Cerebellar Nuclei* (DCN[2]) [9]. The cortex consists of two key cell types, the *granule cells* and *Purkinje cells*, and is organised into three layers, the *granular layer*, the *Purkinje cell layer* and the *molecular layer* (Figure 6).

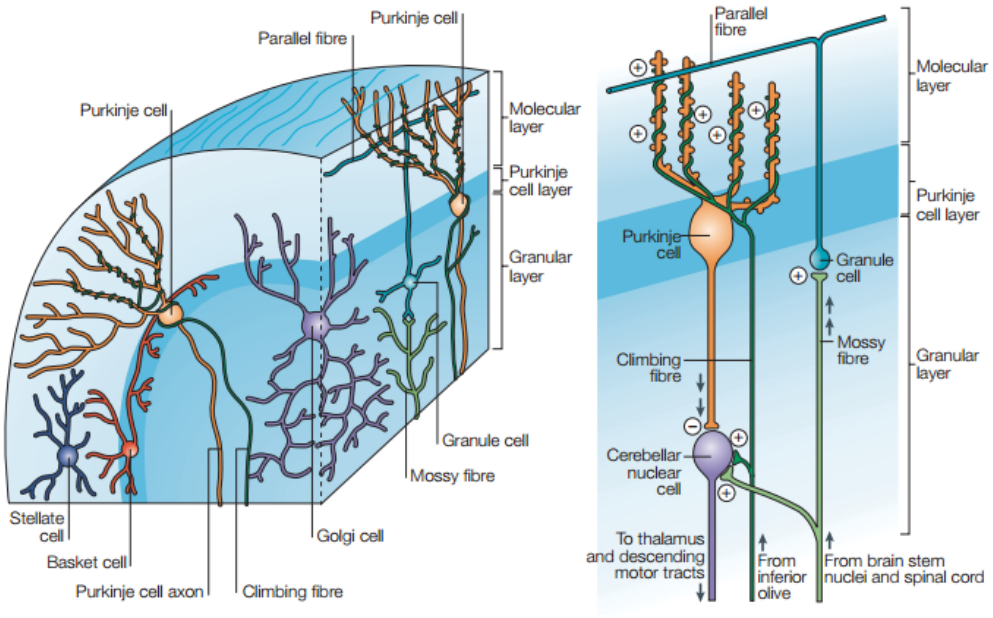


Figure 6. Basic structure of the cerebellar cortex. Reprinted by permission from Springer Nature: Nature, Nature Reviews Neuroscience [73] © 2005.

[2] DCN is used elsewhere in the thesis with different meaning

The granular layer consists mainly of granule cells, which receive input from *mossy fibres*, which in turn provide one of two main afferent pathways to the cerebellum. Granule cells in fact form the bulk of the cells within the cerebellum. There are a variety of sources of this input, including cerebral and sensory, and the latter is the most significant for this thesis. Purkinje cells form their own layer between the granular layer and the molecular layer. Axons of the granule cells form *parallel fibres* which synapse onto the Purkinje cells, and make up the molecular layer. Parallel fibre input to the Purkinje cell is excitatory, but it also receives inhibitory input from *basket* and *stellate cells*, which are both interneurons. The molecular layer also includes *Stellate cells* which receive input from the parallel fibres, synapsing onto the Purkinje cells. Basket cells are found within the Purkinje cell layer and form an inhibitory loop from output to input of the Purkinje cell. Granule cells also receive inhibitory input from *Golgi cells* that are found in the granular layer and receive input from both mossy and parallel fibres, and this may contribute to the richness of the analysis and filtering of the sensory input to the cerebellum as discussed later in Section 3.5.

The second main afferent pathway is the *climbing fibres* that also synapse onto the Purkinje cells. Climbing fibres are the axons of cells that appear in the *inferior olive*, which in turn appear to receive input from the *superior colliculus* [74]. The firing rate of the climbing fibres is orders of magnitude lower than that of the Purkinje cells so that it has no direct influence on the sensory signal yet does influence the weights of the parallel fibre-Purkinje cell synapses [75]. There are a large number of granule cells for each Purkinje cell, on the order of 10^5. Axons of the Purkinje cells carry output from the cerebellar cortex into the DCN[3], and in fact, this is the only output pathway from the cerebellar cortex. Neurons in the DCN form the actual output from the cerebellum. These neurons are inhibited by the Purkinje cells, and also receive input directly from the mossy and climbing fibres, which is excitatory.

3.2.1 The "cerebellar chip"

The notion of the so-called *cerebellar chip* has arisen from the evidence that the regular structure of the cerebellum consists of a large number of microzones of identical "circuitry" whose function depends on external connectivity [8]. The analogy is one of

[3] Deep Cerebellar Nuclei. DCN is used elsewhere in the thesis with different meaning

an integrated circuit (or “chip”, to use the popular terminology), consisting of large numbers of identical circuits. Treating the cerebellum as a single, large NN would mean overlooking the diverse range of functions that are carried out. The smallest possible functional unit might be based around a single Purkinje cell and its associated inputs formed by mossy fibres/granule cells along with climbing fibre input, although the microzone in the mammalian cerebellum is taken to be a group of Purkinje cells that corresponds to an identifiable function. One such possible function is akin to an adaptive filter, which is described in Section 3.5, and which operates on inputs, carrying out some sort of processing, to produce an output. The relationship between input and output characterizes the function of the microzone.

As mentioned in Section 3.2, there are both excitatory and inhibitory cells in the input sections of the “chip”. The role of inhibitory cells may be to act like the inverting input to a differential amplifier; differential inputs to the cerebellum thus transmit the salient sensory input and reject the common-mode noise signal [9].

3.3 Subsumption architecture and the cerebellum

Subsumption architecture is a hierarchical robotics control paradigm in which higher level behaviours are divided in to lower level sub-behaviours (Figure 7), referred to as modules and described as *levels of competence* [76]. Examples of levels of competence from [76] include obstacle avoidance at the lowest level, aimless wandering at the next highest level and exploring at the level above that, with each layer utilising the competence below it. This approach allows the development of a robot system from the bottom-up, starting with low level competences and adding higher levels as more sophisticated behaviour is required. Input is typically from sensors and output is to actuators or is used in the suppression or inhibition of Input/Output (I/O) to lower layers. Through this I/O manipulation higher level competences can subsume lower level competences. An advantage claimed of this approach is increased robustness in that if a high-level competence cannot react quickly enough to a changing situation, it will then fail to suppress its lower level competence which can then resume more control (the assumption of course is that the lower level functions can operate more efficiently or with greater speed, which is not unreasonable).

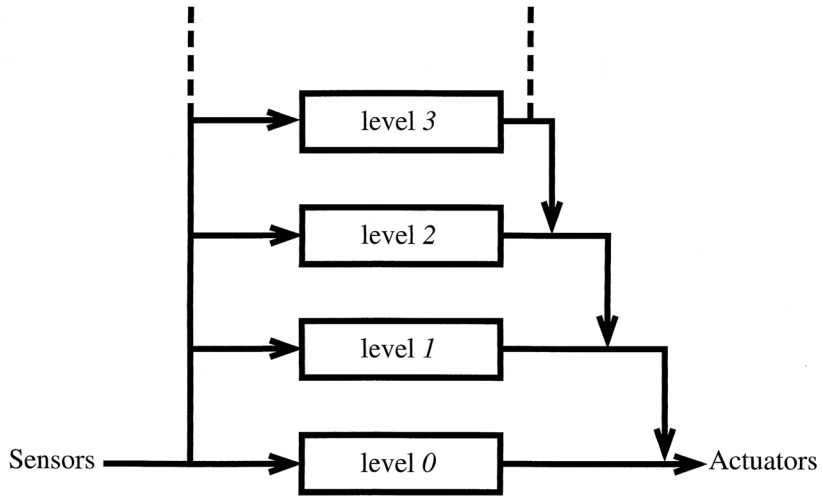


Figure 7. Hierarchical layers in subsumption architecture. Reprinted with permission from [77]. © 1999 Massachusetts Institute of Technology, published by the MIT Press.

Montgomery and Bodznick [9] suggest that the role of the cerebellum could be akin to a higher level module that sits over lower level functions of the brain and subsumes those functions as required by the task at hand, providing short term adaptation until the lower level function is able to adapt over a longer term, when the cerebellum can then "step back" and allow the lower level module to regain more control. This does not seem to be fully true to the subsumption architecture in that the latter builds higher level competences that remain in operation permanently, however, it is an interesting and informing analogy nonetheless. The analogy as it applies to this thesis would be that the SSL unit acts as a low-level competence (one could extend the analogy and think of further, lower level competences such as the audio capture subsystem), which can provide an accurate azimuth estimate under normal circumstances, that is, in an anechoic environment free from interfering noise sources, but which may produce erroneous estimates in non-ideal environments. The cerebellar calibrator described in this thesis sits on top of this and subsumes the SSL function by interjecting a calibrating signal. This very nicely reflects the subsumption architecture, in which low level modules are unaware (so to speak) of the interference of the high-level module, which is the case here as the calibration signal is added to the output of the SSL unit. In other words, the calibrator does not alter the operation of the SSL unit itself, but interjects a calibrating signal at the output.

3.4 The role of the cerebellum in cognition

Montgomery and Bodznick [9] provide a very nice review of the role of the cerebellum in providing a sense of self and agency. This is facilitated in part through the function of the cerebellum as an adaptive filter (see Section 3.5) that monitors the efferent (outgoing) signals from the brain (e.g. motor commands) along with the re-afferent (resultant incoming sensory) signals, and where they correlate, cancelling the resulting sensation that would otherwise be perceived as being generated by a different agent [78]. There has been recent recognition that the cerebellum plays a role in perceptual processes [79], with auditory input to the cerebellum as well as other sensory input such as vision. Much of the research concerning the role of the auditory cerebellum however seems to be in the area of speech production. This makes sense given the cerebellum's historically perceived role in motor control as speech production is partly a motor problem.

3.4.1 The cerebellum and cerebellar-like structures in auditory processing

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 that it plays a role in perceptual processes [9, 79, 80]. The role of the cerebellum in auditory processing in particular was recognised several decades ago [81], although it is only relatively recently that this aspect of cerebellar function has received much attention [82-86]. Work in this area has mainly focused on speech perception and production- there is very little on the role of the cerebellum in auditory localisation. The DCN[4] appears early in the auditory pathway (Figure 2) and has a cerebellum-like structure that takes part in the spectral localisation of higher frequency audio using ILD. There is no mention in the literature of the role of the cerebellum in ITD, however, it is an intriguing thought that there might be such a role given the universality of the "cerebellar chip" and the role of the cerebellum in temporal aspects of auditory processing [87, 88] and event timing [89], albeit on timescales much larger than those involved in ITD. As such, there is currently no biological plausibility for the approach taken in this thesis, and it can as yet only claim a new demonstration of the universal utility of the "cerebellar chip" in robotics.

[4] Here, DCN is Dorsal Cochlear Nucleus

An intriguing insight into the role of the cerebellum in auditory processing, but at a high level, interfacing with conscious thought, is the differentiation between external sounds and those generated internally through a person's thoughts. In the same way that the cerebellum has been shown to have a role in cancelling the self-generated tickle sensation, it has been shown that patients who suffer from episodes of hallucinatory audio also suffer from an inability to distinguish self-generated tactile sensations [78, 90].

3.5 Adaptive filter model of the cerebellum

3.5.1 Introduction

The adaptive filter model of the cerebellum has been used successfully in studies including a variety of robotics applications, although these are quite limited in number, despite it being a powerful and versatile algorithm. These have included [4, 5, 91-93].

A general filter is a system that receives a time varying signal as input, applies a transform and produces a time varying output that depends on the input in a way that is characterised by the transform, and as the name suggests, in so doing usually removes an undesirable feature of the input such as unwanted noise. The transform is usually characterised by a set of parameters (these could be, for example, the coefficients of a difference equation in a discrete-time filter, where the output of the filter depends on the weighted sum of past and present inputs), and the filter design problem is one of arriving at a set of parameters to achieve the desired filter transform.

An adaptive filter is one which has adjustable parameters, whose values are determined through some algorithm that typically is designed to optimise the performance of the filter, often by minimising performance error. An example might be the *Least Mean Square* (LMS) algorithm [94], which is a *gradient descent* algorithm intended to optimise filter coefficients by iteratively finding the gradient of the *Mean Squared Error* (MSE) to minimise the same. Adaptive filters have applications in engineering where the desired characteristics of the filter cannot be known *a-priori*, for example in acoustic noise cancellation, where the nature of the unwanted signals to be removed will depend on the acoustics of the environment and the nature of the sources of noise.

The adaptive filter model of the cerebellum was proposed by Fujita [71] as a variation on the Marr-Albus model [95, 96]. This model emphasises the resemblance of the cerebellar microcircuit to an adaptive filter [8, 72, 97, 98]. The cerebellar microcircuit and the adaptive filter model are shown in Figure 8 A and B respectively. The model is characterised by a very rich set of inputs/basis filters [9] (note, however, that the models in this thesis do not have basis filters). This richness is a result both of the very large number of granule cells (and hence parallel fibres) and of additional processing carried out within the granule cell network itself, which is enhanced by Golgi cells in the granular layer (Section 3.2), and contributes to the power of the adaptive filter function by providing a massive signal analysis capability. The large number of mossy fibres also allows input to the cerebellum from very diverse areas of the brain and sensory systems. Sensory input is to granule cells via the mossy fibres. Granule cell axons form parallel fibre inputs to Purkinje cells. The granule cells (along with their parallel fibre axons) form basis functions/filters that may include, for example, delay functions. In this thesis, and the precursory work that the cerebellar calibrators were based on [4], basis filters were simply unity gains. Hence, the (mossy fibre) input is analysed into multiple filter pathways and synthesized at the Purkinje cell with weights that are affected by the climbing fibre input to the Purkinje cell. Whereas the parallel fibres convey sensory input signals, the climbing fibre conveys a teaching signal. Learning is based on the covariance learning rule [8, 9, 72, 97-101]. This rule serves to de-correlate input signals that contribute to error. As Purkinje cell output is inhibitory, parallel fibres that correlate with Purkinje cell activation will have their synaptic weights reduced by the learning rule. Purkinje cell activation coincides with error (e.g. retinal slip in the vestibulo–ocular reflex) and de-correlation learning drives Purkinje cell activity towards zero. Activation of the Purkinje cell has an inhibitory effect on the appropriate deep cerebellar neurons, counteracting the error. Those parallel fibres that are active in a way that is uncorrelated with Purkinje cell activation will have their weights increased. If a parallel fibre is inactive, the covariance learning rule means that its synaptic weights will be unchanged.

3.5.2 Implementation of the adaptive filter model of the cerebellum

In this thesis Matlab was used to implement an adapted version of the model used in the Bella project (Section 1.1) on a desktop PC/laptop, and this is described in detail in

Chapter 4. As mentioned in that chapter, a Zynq *System on Chip* (SoC) could be a potential candidate for in-the-field implementation. More details are given in Section 4.2. As mentioned in Section 2.2.1, *neuromorphic engineering* provides a potential low power consumption solution and examples exist of neuromorphic implementations of the cerebellum [102, 103] and specifically the adaptive filter model of the cerebellum [92].

3.6 The cerebellum in motor control

The motor control role of the cerebellum may not seem relevant to his thesis, however, as the thesis "borrows" a multiple models approach to selecting models for motor control, described in Chapter 5, a brief treatment of this role is undertaken here.

Models of the cerebellum have increasingly been used as adaptive controllers [104]. Internal models contain a paired forward and an inverse model. In the context of motor control, input to the forward model is the motor command but could also include sensory input [74] and the output is a prediction of the consequences of the action given the current state and motor command as input as the forward model learns the dynamics of the controlled plant [6, 105]. In this way, the forward model overcomes the problem of large delays in sensory feedback. On the other hand, an inverse model of the system acts as a controller, providing feedforward control [104] as it produces the required motor command for a given desired state. Another way of viewing this is that the inverse model acting in series with the controlled plant together form an identity such that the output is identical to the input provided the inverse model is accurate.

Forward models can be learned through experience, adjusting for example synaptic weights in response to an error signal obtained from the difference between desired state and actual state ascertained through feedback.

Inverse (controller) models are more problematic to learn, as the teaching signal is not usually directly available. Two key solutions to this have been proposed: distal supervised learning [106] and feedback-error-learning [107, 108]. In the latter, the feedback error signal is provided in motor coordinates (in the cerebellum this is via the climbing fibre inputs). In this scheme, the feedback error signal is in task-oriented coordinates (e.g. desired arm trajectory) or body-oriented coordinates and is transformed into motor coordinates by an approximate inverse model of the plant, which is used to teach a feed-

forward adaptive element such as a cerebellar model. Porrill et al. eliminated the need for such an approximate inverse model by applying the teaching signal instead to a recurrent adaptive element [7].

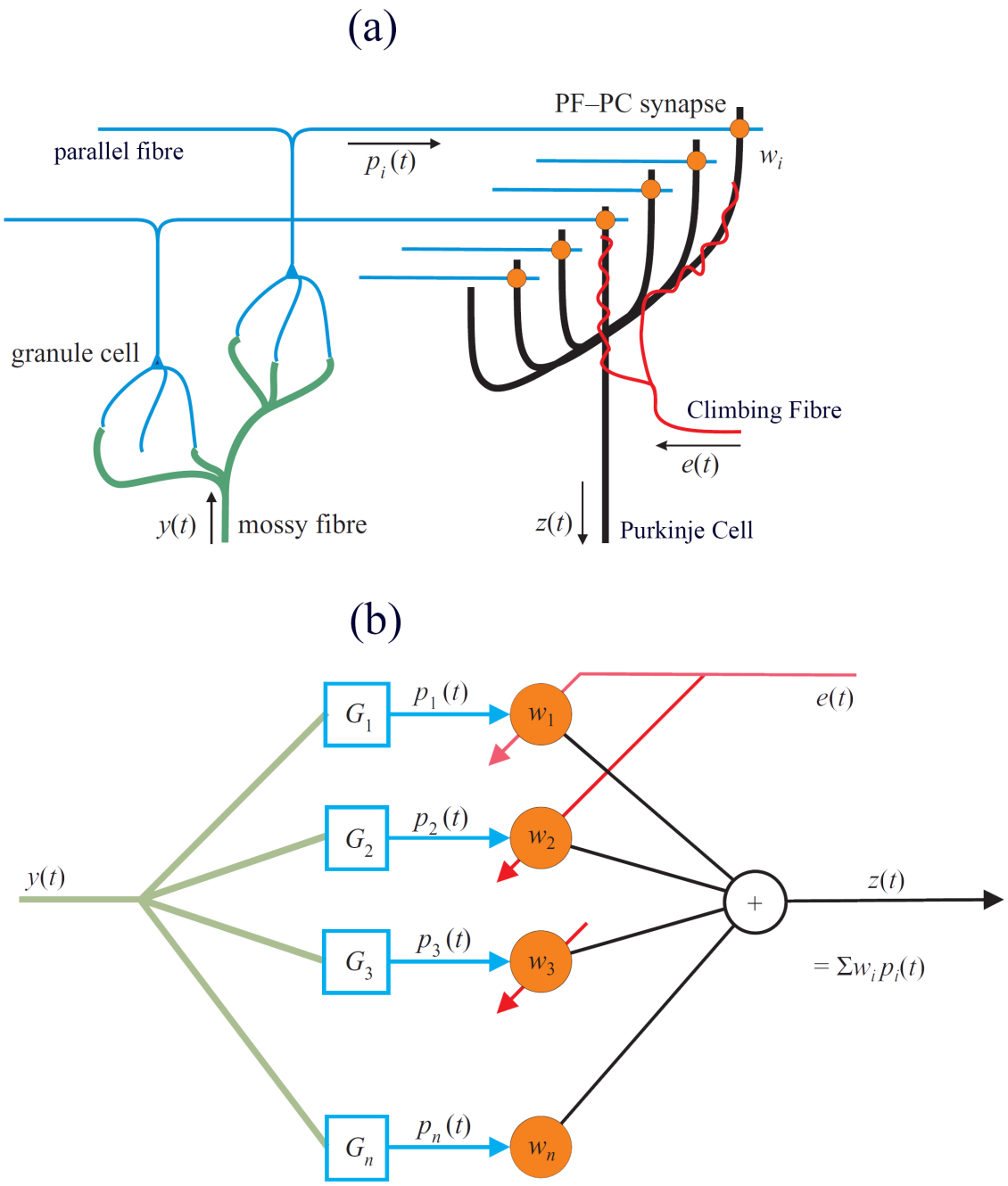


Figure 8. Adaptive filter model of the cerebellum. (a) Cerebellar microcircuit. (b) Adaptive filter equivalent. Adapted with permission by Royal Society from [7].

Chapter 4 Cerebellar calibration

4.1 Introduction

4.1.1 Adapted model- whisker map calibration

The cerebellar calibration model is an adaptation of that used in a previous study to calibrate whisker input to a robot platform [4, 5], which draws on the adaptive filter model of the cerebellum to calibrate a 2-dimensional topographic map of the whisker sensory space [71]. That study also drew on the concept that animal collicular maps receive information about target locations from multiple modalities [109], and that saccades in particular can be controlled based on sensory input to, for example, bring the target onto the fovea in primates. Accuracy can be relearnt following artificial mis-calibration where the Cerebellum is intact, suggesting a role for the latter in calibration of collicular maps [110]. An array of 20 whisker-like tactile sensors were mounted on the "head" of a robotic manipulator, with a camera mounted centrally to the whisker array (Figure 9). The model was implemented in Matlab and cerebellar connections are shown in Figure 10, with the topographic map representing the superior colliculus. Input to the model were coordinates of the sensed whisker contact location in head-centric space, and represented by a 2-dimensional Guassian on the topographic map. Granular layer processing is limited to coarse coding (with 64 parallel fibres in an 8 x 8 2D array) and normalisation, so that there is no explicit basis filter. The performance of the system using artificial distortion (in software) is shown in Figure 11.

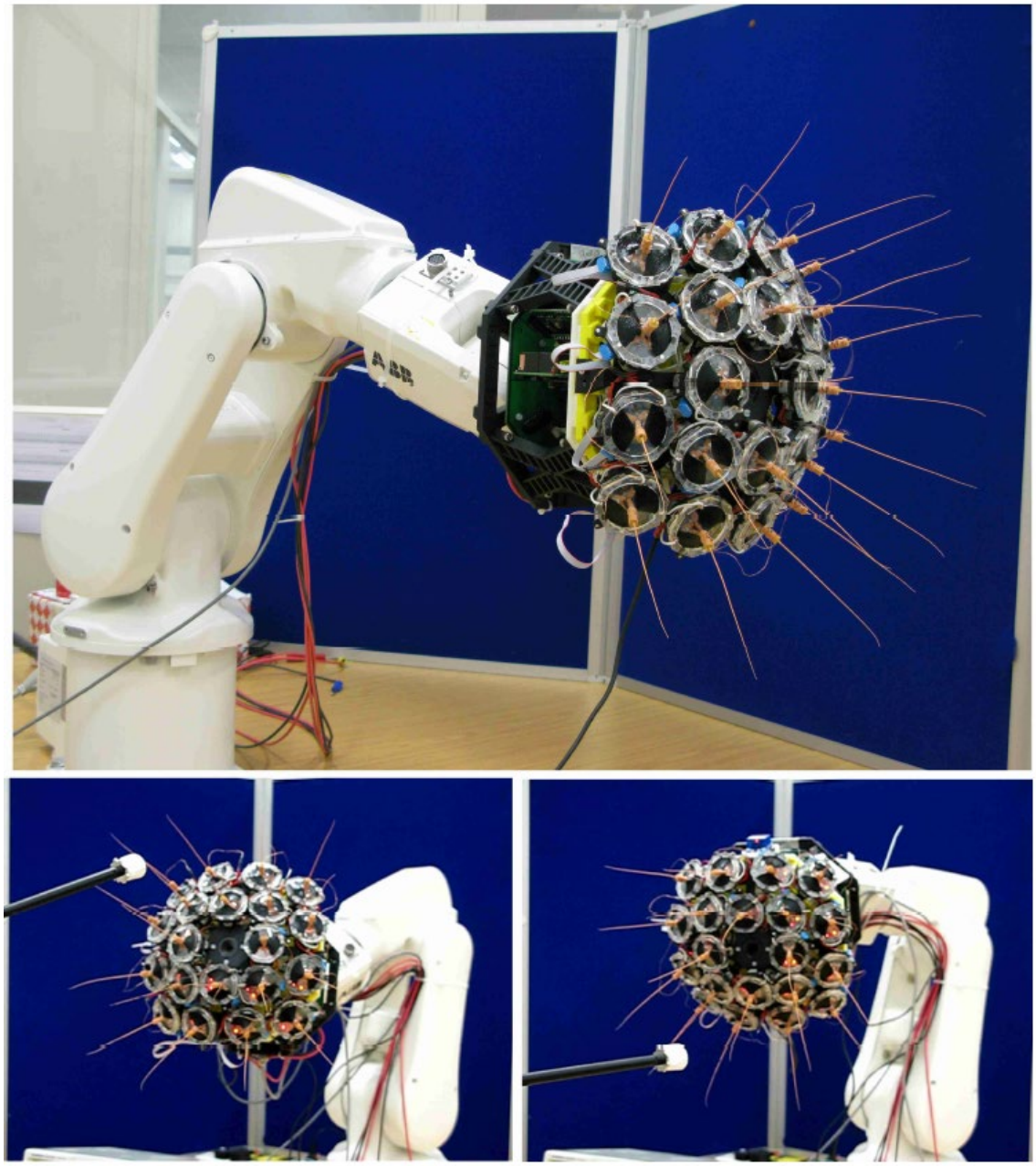


Figure 9. Bellabot platform. Reprinted from [4]. © 2016 IEEE.

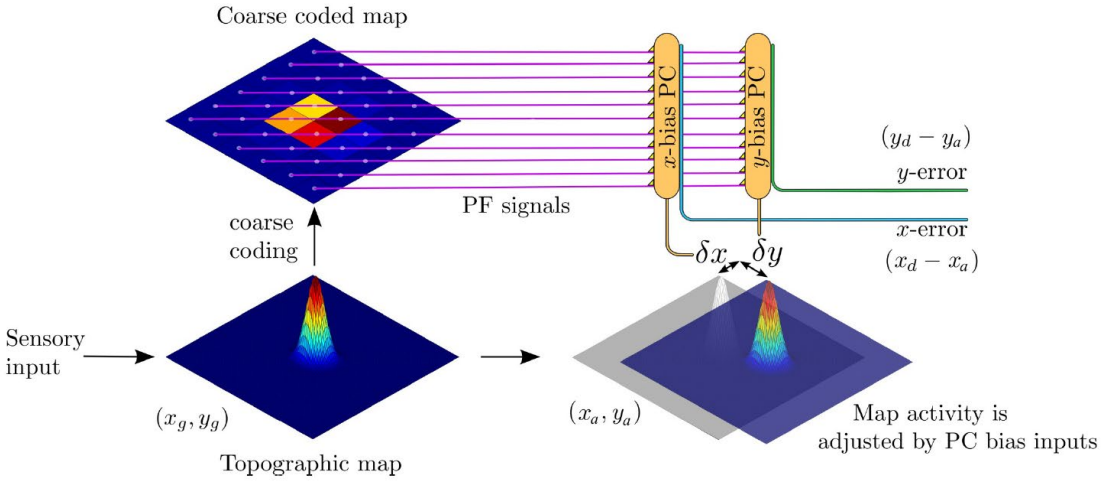


Figure 10. Architecture of the whisker map calibration system. Unimodal sensory signals, corresponding to target locations (xd, yd), are written into the map, which because of the distortion provides an inaccurate estimate of target location (xg, yg) that is used to generate a correspondingly inaccurate orienting response. The map output is also sent to the cerebellum, where it is converted into a coarse-coded, normalised set of PF signals. These are sent to two cerebellar microzones, each represented in the diagram by a single Purkinje cell, that receive climbing fibre inputs that initially signal errors (xd-xg, yd-yg) in the orienting response. These errors are used to alter PF-PC synapses, generating cerebellar output that shifts the map so that the orienting response is now made to the new location (xa, ya). This process is repeated until the error (xd−xa, yd-ya) becomes zero. Figure and caption adapted from [5] under the Creative Commons Attribution License.

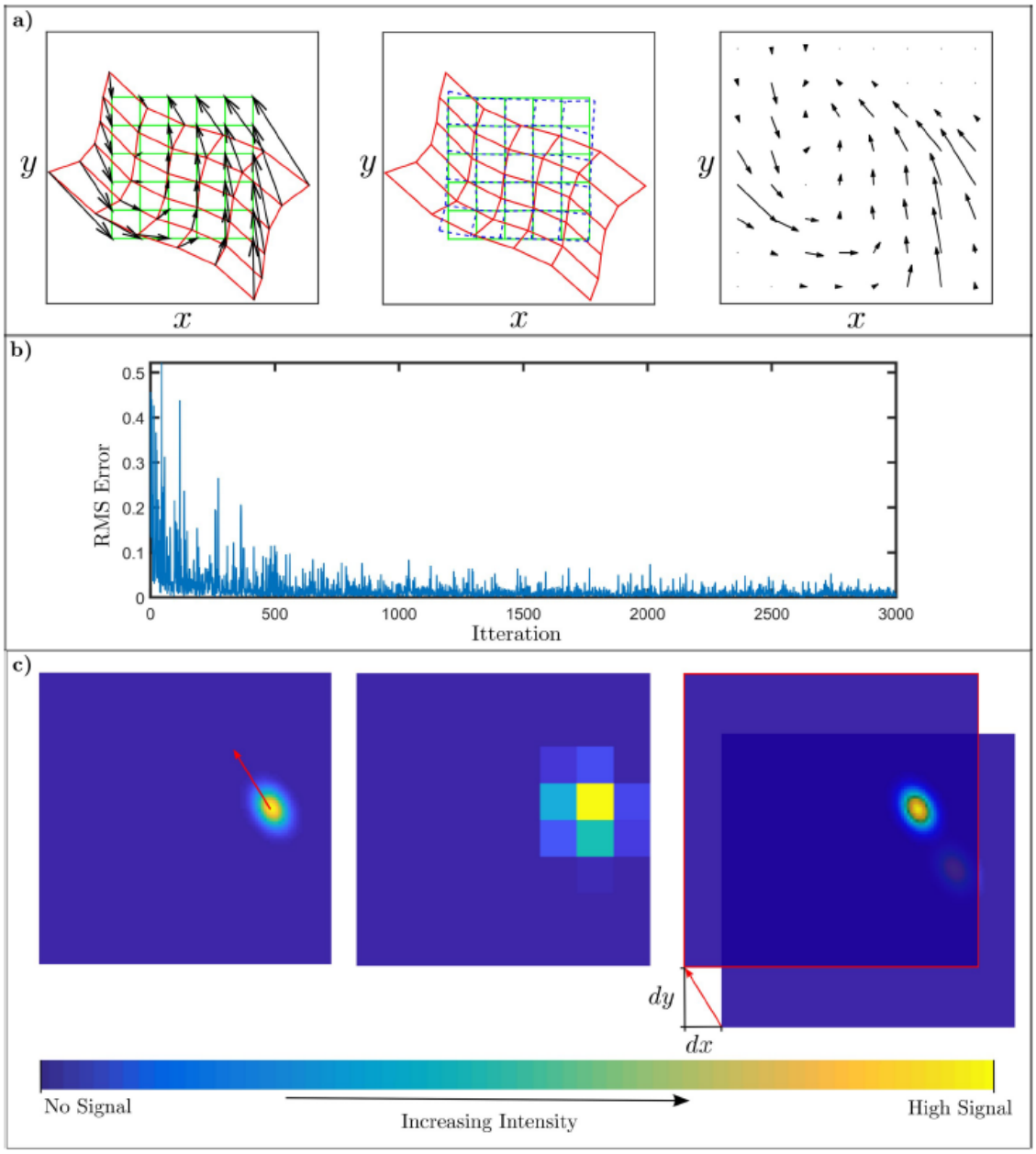


Figure 11. Recalibration of a single target map with curvilinear distortion. a) The left hand panel shows an initially accurate map (green line) in the superior colliculus (SC), with artificially induced curvilinear distortion (red line) (details in Methods). The shifts in the map to correct for the distortion are dependent on the location in the map, and are indicated by black arrows. The learnt cerebellar recalibration of the distorted grid (teal line) is shown in the middle panel. The right hand panel shows the combined learnt weights in the x- and y-directions corresponding to each coarse coded parallel fibre signal (weights initially zero). b) Time course of recalibration, showing how RMS errors in orienting responses change with number of target presentations. c) Example of learnt dynamic cerebellar recalibration. The left-hand panel shows the shift in the map

(red arrow) required to produce an accurate orienting response to the inaccurate target location provided by the distorted map. The centre panel shows the coarse-coded, normalised parallel fibre signals produced by the inaccurate target location. The right hand panel shows that after learning the parallel fibre signals now shift the map by just the required amount to produce an accurate response. Figure and caption reprinted from [5] under the Creative Commons Attribution License.

The Purkinje cell in the model synthesizes the parallel fibre signals modulated by the synaptic weights into a collicular x-y map-shift shift signal that is applied as a bias to the output from the whisker array. The amount of bias is the weighted sum of the parallel fibre inputs:

$$\delta x = \sum_{i=0}^{n} w_{xi} p_i \tag{3}$$

$$\delta y = \sum_{i=0}^{n} w_{yi} p_i \tag{4}$$

Where w_x and w_y are the parallel fibre-Purkinje cell weights associated with the x and y direction respectively. Weight updates are based on the covariance learning rule [8, 9, 72, 97-101]:

$$\Delta w_{xi} = -\beta e_x p_i \tag{5}$$

$$\Delta w_{yi} = -\beta e_y p_i \tag{6}$$

Where β is the learning rate, e_x and e_y are the x and y target position errors respectively.

A limitation of the implementation in [4] and [5] is that there is no stopping criteria during learning. The number of weight update trials used in [4] is not discussed in that paper but from the results presented appears to be between 110 and 140 trials. The number of trials used in [5] was 3000 although no rationale is given for the choice of this figure.

Another limitation of this implementation is that the compensating shift produced by the cerebellar circuit is a global map shift, based on the simplification that only a single target is considered. Wilson et al. assert that it is likely that separate cerebellar microzones would be required to calibrate different areas of the collicular map should multiple targets be used.

4.1.2 Adaptation to SSL calibration

The software that was adapted for this thesis was not that used on the Bellabot but rather a demonstration version, which used artificially generated whisker output data in software (the Bellabot software itself would have been too unwieldy for the purposes of the thesis, consisting mainly of control and communication routines specific to the Bellabot platform). Otherwise, the calibration model is the same as that described in section 4.1.1, with some modifications that are described in this section.

The adapted system is shown in Figure 12, adapted for audio input. The key changes made to that software are that it was adapted to represent a 1-dimensional case and with inputs that are generated from the SSL input rather than whisker input, otherwise the model remained largely unchanged. This was done as, at the time, it was envisaged that the system proposed in the thesis could form another modality in the Bellabot system itself, so it was considered prudent to adapt the software as little as possible. Input was the SSL estimate (azimuth) transformed into distance from centre (directly ahead of the robot head) along a circa 2.5m arc (see below), so that the inputs were between 0 and +/- 1.25m (representing an azimuth of +/- 90º, although inputs were restricted to +/-45º). The map stores a probabilistic representation of sound source azimuth in robot head-centric space, but using a 1-dimensional Guassian rather than 2-dimensional as in Bellabot. Otherwise, the Guassian parameters remained the same. A course-coded version of the map, as with Bellabot, transmits activity at each place on the map to the cerebellum via the parallel fibres. The weights w_i of the parallel fibre-Purkinje cell synapses are updated in the same way as in Bellabot by the covariance learning rule.

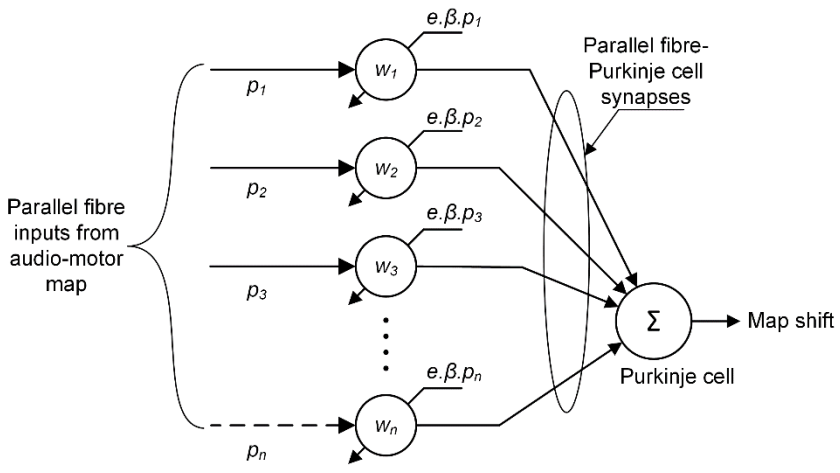


Figure 12. Cerebellar calibration using the adaptive filter model of the cerebellum. Reprinted from [54]. © 2018 IEEE.

An audio stimulus results in activation of the SSL unit, which provides an estimate of the azimuthal position of the sound source. This is represented internally in the robot system as the length of the arc from the centre position of a circle that encompasses and is centred on the robot head to the point on that circle occupied by the sound source (Figure 13). This was done rather than converting to an azimuth internally in order to stay close to the implementation of the system in [4]. The conversion from azimuth to arc length is given by

$$\frac{2\pi d\theta}{360} \tag{7}$$

where d is the distance from the robot head to the sound source and θ is the azimuth of the sound source in robot head-centric space. It is the length of the arc from the centre position that is input to the cerebellar model.

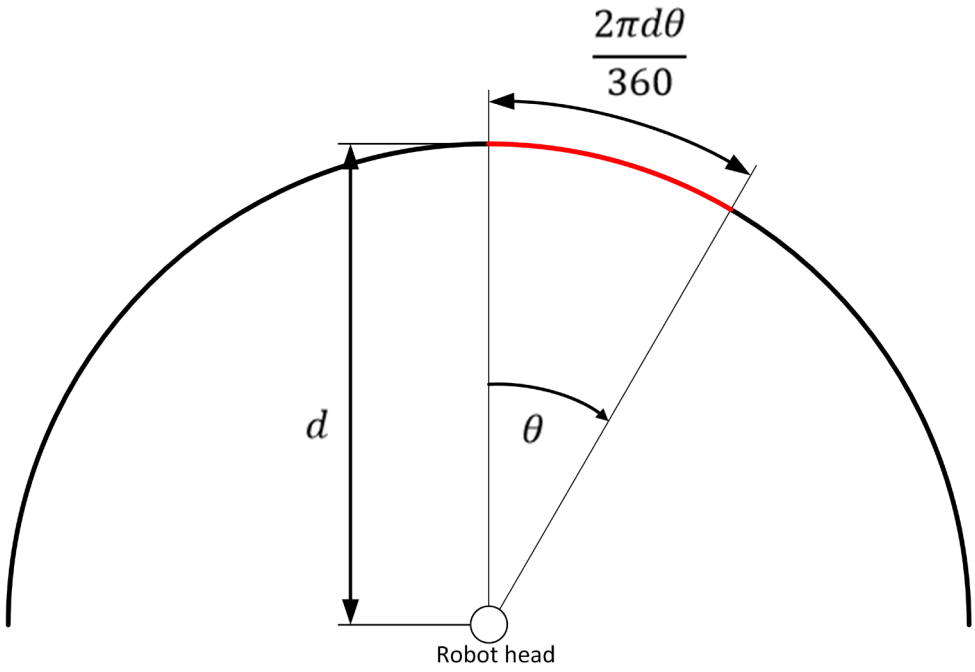


Figure 13. Internal representation of azimuth.

This internal representation is divided into a regular grid with activity in each cell of the grid forming one input (i.e. the mossy fibre/parallel fibre) into the cerebellar model, and is course-coded for computational convenience as described in [4], so that the activity on a single parallel fibre is a probabilistic representation of a number of azimuthal positions (in this case the azimuth range is divided into 8 regions (with 8 parallel fibres), following the approach in [4], each representing approximately 10° segments). For the 1-

dimensional SSL work carried out in this thesis, this course coding may not have been necessary, but could become so if the system is extended to two dimensions (e.g. elevation as well as azimuth) or more (e.g. distance to sound source). The Purkinje cell, represented by the summing element in Figure 12, synthesizes the parallel fibre signals modulated by the synaptic weights into a (positive- toward the right, or negative- toward the left) azimuth shift signal that is applied as a bias to the output from the SSL unit. The amount of bias is the weighted sum of the parallel fibre inputs, adapted from equation (3):

$$\delta\theta = \sum_{i=0}^{n} w_i p_i \tag{8}$$

where n is the number of parallel fibres, w_i is the i_{th} synaptic weight and p_i is the activity on the i_{th} parallel fibre.

The weights w_i of the parallel fibre-Purkinje cell synapses, initially zero, are learned using the covariance learning rule [111], equivalent to the LMS algorithm in an adaptive filter, and updated as in [4], adapted from equation (5):

$$\Delta w_i = -\beta e p_i \tag{9}$$

where β is the learning rate, p_i the activity on each parallel fibre and e is the orient error, that is, the difference between the ground truth azimuth of the sound source and the calibrated SSL output (in fact, arc length is used rather than azimuth, as mentioned above).

Figure 14 shows plots of parallel fibre activity, at 0° ground truth azimuth, for two different models which have learned with an azimuth error applied within the algorithm (by simply adding a constant shift to the SSL estimate). Line charts have been used for clarity although the functions plotted are not continuous, but rather are discrete with integer index values. The blue curve represents the weights of a model that has learned with an azimuth error of -20° (somewhat exaggerated errors were introduced for the purposes of demonstrating the effect), while the orange curve shows the weights of a different model that has learned with a +20° error. This is at 0° ground truth azimuth (that is, with the sound source located dead ahead of the robot head). The orange curve is for an azimuth error of +20° and the blue curve is for a -20° azimuth error. Lower parallel fibre indices represent activity to the left and higher values to the right (index values 4 and 5 span the 0o ground truth azimuth position). The shape of the curves reflects the

Gaussian function that has been applied to the parallel fibre inputs as described above. Higher value parallel fibre indices represent (a set of) azimuths toward the right and lower value indices toward the left with respect to the centre position which is directly ahead of the robot.

Figure 15 shows plots of cerebellar weights for the same models whose parallel fibre signals are shown in Figure 14. Again, line charts have been used for clarity although the functions plotted are not continuous, but rather are discrete with integer index values. The orange curve is for an azimuth error of +20º and the blue curve is for a -20º azimuth error. Lower value weight indices represent Purkinje cell synapses with input corresponding to azimuth positions toward the left and higher values toward the right (index values 4 and 5 span the 0º azimuth position). It is clear that the models have learned to calibrate in opposite directions (for example, the model that has learned with a +20º azimuth error produces a negative compensatory bias).

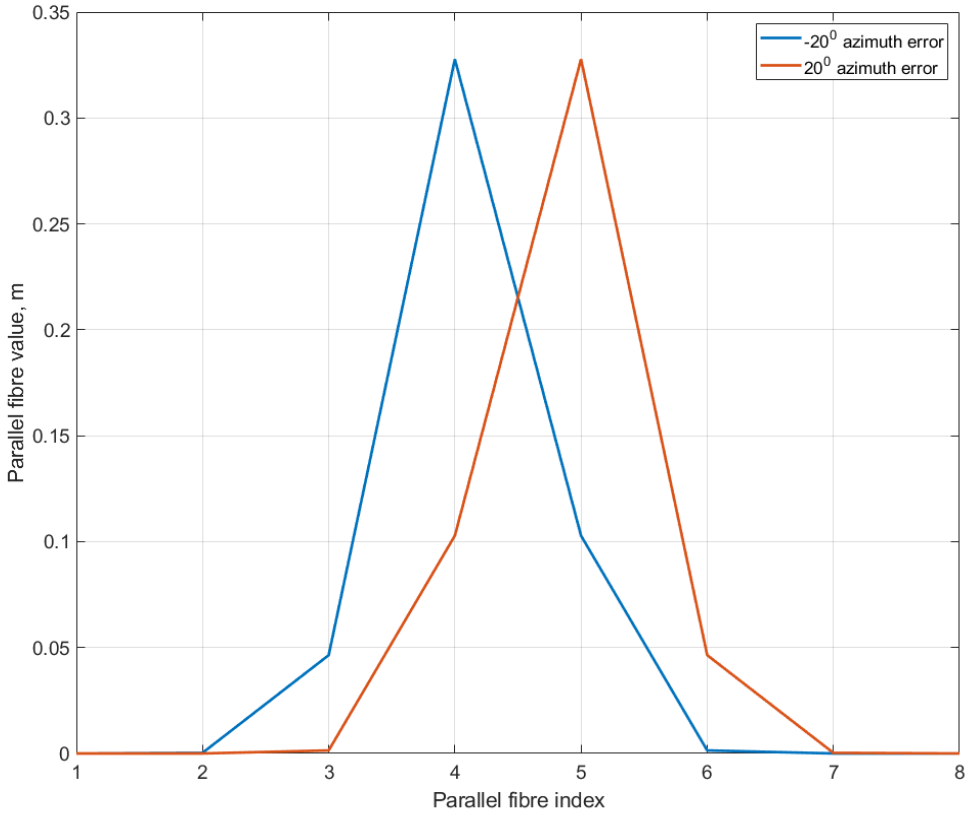


Figure 14. Parallel fibre activity for two different azimuth errors.

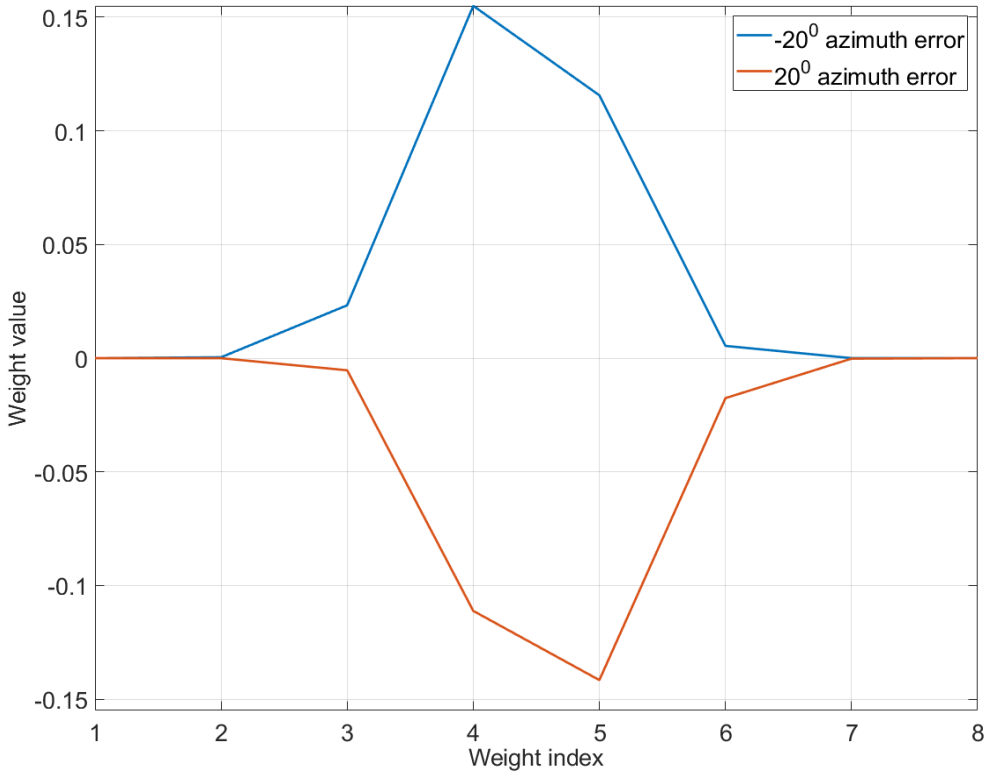


Figure 15. Cerebellar weights for two different azimuth errors. Weight values are dimensionless.

The calibration is carried out in one plane only, that is, it is based on calibrating a 1-dimensional single direction of arrival SSL paradigm as categorised by [112]. The system could be extended to two-dimensional SSL (azimuth and elevation); indeed, in terms of the cerebellar calibration this would be quite straightforward, as the work in [4] is a two dimensional whisker map. Achieving this with the basic SSL technique used in this thesis, however, might be problematic if a binaural approach is to be favoured. Either microphone arrays of at least three (preferably 4) microphones would be required to do so based on ITD, or an asymmetrical head/pinnae arrangement would be needed to do so with ILD.

The cerebellar model is shown in situ in Figure 16. In the full system, the calibrated output from the audio-motor map is used to orient the robot head toward the sound source, and a visually derived error after orientation is used as a teaching signal to adjust the weights of parallel fibre/Purkinje cell synapses, which are initially set at zero, although, as described in Section 4.2.2, the odometry of the experimental apparatus was used in this

thesis. Post learning, the cerebellar model applies a shift, according to Equation (8) to compensate for errors in the SSL algorithm output. Through repeated updating of the cerebellar weights according to Equation (9), with the SSL algorithm presented with sound from sources of randomly selected azimuth, the cerebellum learns to compensate for errors introduced into the SSL estimate. As described in Section 4.2, this was done by placing the sound source at random azimuths with respect to the robot head. Of course, this raises the question of how this learning would take place in the field, and the assumption is that the robot would move around in a random fashion to achieve the same effect.

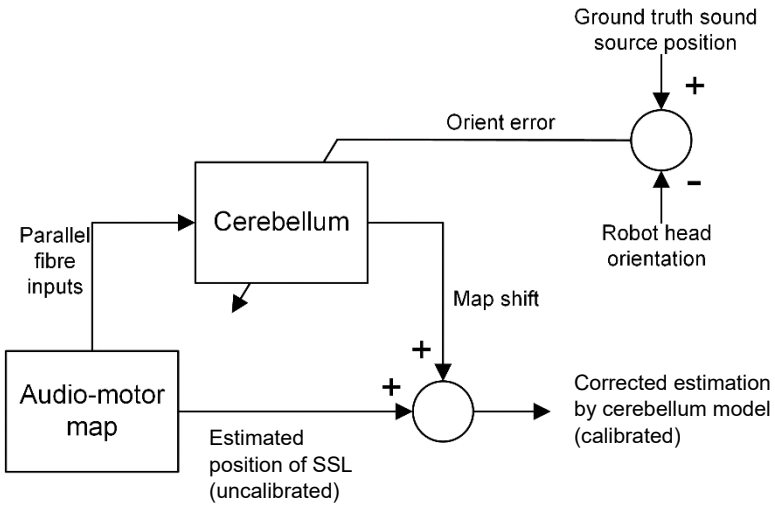


Figure 16. Cerebellar model in learning mode. Adapted by permission from Springer Nature [113] © 2017.

4.2 Method

Matlab was used to control experiments and for implementation of algorithms, on a desk based PC/laptop, which were connected to various apparatus described below.

Although not ultimately used, a gammatone filter bank (Section 2.2.1) was implemented on the Zedboard Zynq SoC development board. The Zynq SoC combines an ARM core and *Field Programmable Gate Array* (FPGA) on the same fabric which can be used independently or together [114]. Because it was not used in the project, it is mentioned only in passing here, however, such a system could be a candidate solution for in-the-field implementation in future work.

Two microphones (Audio-Technica ATR-3350 omnidirectional condenser lavalier) were mounted in free field at the extremities of a horizontal bar (Fig. 6), with an inter-

microphone distance of 0.16m and were connected to a computer using a M-Audio MobilePre USB audio capture unit with a sampling rate of 44100Hz. Sound pressure level at the microphones was measured using a Max Measure MM-SMB01 sound level meter and was maintained at approximately 70dBA with the sound source directly facing the robot head.

A low-cost USB Webcam (TeckNet) was mounted on the robot head centrally between the microphones with a view to using visual sensory feedback to ascertain the ground truth location of the sound source position in future work. This was not used in the thesis, as explained in Section 4.2.2, apart from a pilot experiment, described in Section 4.3.3.

A sound source (Logitech Z150 Speaker) was positioned at a fixed distance from the robot head (Figure 17) and was connected to the computer sound card. A 1 second duration Gaussian noise signal was fed to the speaker to generate audio stimuli (with the exception of one experiment, where a 500Hz pure tone was used- see also Section 4.3.4).

For the work described in this chapter, the sound source position itself was fixed, and the robot head was instead rotated to generate stimulus from various azimuths. This is because the sound source would need to be placed manually which does not easily allow the conduct of reproducible trials with many data points. This seemed a reasonable initial approach, allowing automated running of experiments, and crucially made a number of exploratory pilot experiments feasible. There are precedents for this approach, e.g. [115], although there are limitations to this technique. The most obvious is that it is not a true reflection of what would happen in a real situation. As the robot head rotates to mimic the displacement of the sound source, the environment "moves" with the source in this arrangement- that is, the sound source's position in the environment would remain unchanged (it would simply change in position with respect to the robot head). Later on, apparatus was constructed to allow the automatic placement of the sound source in the environment, and this is described in Chapter 6, and was itself replaced by a re-designed apparatus described in Chapter 7.

The robot head was placed on a commercial Pan-and-Tilt Unit (PTU), an eMotimo TB3. The PTU was connected via USB to the computer and controlled using Matlab. The apparatus is shown in Figure 17.

For all of the SLL calibration model experiments, the learning rate β was fixed at unity as it was in the software provided by the Bellabot project (although the value is not

mentioned in [4] and [5]). Different values were required for satisfactory functioning of the RP models, and these are described in the relevant sections.

4.2.1 SSL error

As the motivation for this study was to calibrate a distorted SSL estimation, various means were used to introduce distortion, beyond that introduced by imperfections in the environment. Simply relying on the acoustic characteristics of the experimental environment was not sufficient- any SSL error was small and not reliably reproducible. In any case, there was a need (in later chapters) to introduce errors for different acoustic environments.

Methods used included artificial distortion post-estimation within the computer algorithms (this was the method used to generate the plots in Figure 14 and Figure 15); moving one or both microphones from the correct position; masking one of the microphones with a physical object. For the work described in subsequent chapters, a means of automatically introducing error in a more reproducible way was developed, and this is described in those chapters where it was used (Chapter 6, Chapter 7, Chapter 8 and Chapter 9). The artificial error in SSL described in this chapter was introduced by multiplying the input to the model (which is a displacement around the semi-circular track) by a constant factor (1.3), which introduces an azimuth-dependent error in the SSl estimate.

4.2.2 Ground truth

Learning of the models requires the availability of the ground truth sound source position, from which an error signal can be derived. In the MOSAIC framework [6] (see Chapter 5), this is provided by sensory feedback, and the work in [4], on which the cerebellar model described in this thesis is based, uses visual feedback to provide the ground truth position of the target. In this thesis, sensory feedback was not used (apart from in one pilot experiment described below and in Section 4.3.3), and the odometry of the experimental apparatus was used instead. This was done for convenience and allowed a wide range of experiments throughout the thesis to be conducted using recorded audio signals. It is assumed that the odometry is sufficiently accurate for this to be treated as providing the ground truth. It is envisaged that a robot operating in the field, however, would use sensory feedback to ascertain the ground truth sound source position, as in MOSAIC, and in [4], and that typically this would be vision (we assume that vision is

good enough to be treated as providing the ground truth). For this reason, the apparatus was designed such that the ground truth could be determined using vision at a later stage. With this setup, the robot head would orient toward the *calibrated* estimated position, to bring the sound source into the field of view of the camera. A pilot experiment was conducted to check that the system using vision would learn in a similar way to that using the odometry for ground truth. As this was not a focus for the thesis, it is reported only in passing here, for completeness. An image from the camera was captured into Matlab (after orientation) and used to ascertain the horizontal position in the image of the sound source. Displacement of a known point on the sound source from the centre of the image could be used to derive an error value. There are a number of ways that this could be done, but for convenience, a light source attached to the sound source was detected visually and transformed into an error signal. Image processing to ascertain the error was carried out in openCV using C++ wrapped in a MEX function in Matlab, rather than using native Matlab functions, to make it easier to transfer the code to a different platform at a later stage should this be required (especially if the system is to be implemented on a mobile robot platform). At this stage, image processing simply finds the horizontal location of the brightest pixel (after blurring of the image to minimise the occurrence of erroneous readings). It is envisaged that rather more sophisticated object recognition may be called for in a real-world scenario. No attempt was made to draw up a trigonometric transform between the image pixel index and the ground truth azimuth; rather, a look-up table was used to map a pixel index output from the image processing algorithm to an azimuth value, with the entries determined empirically.

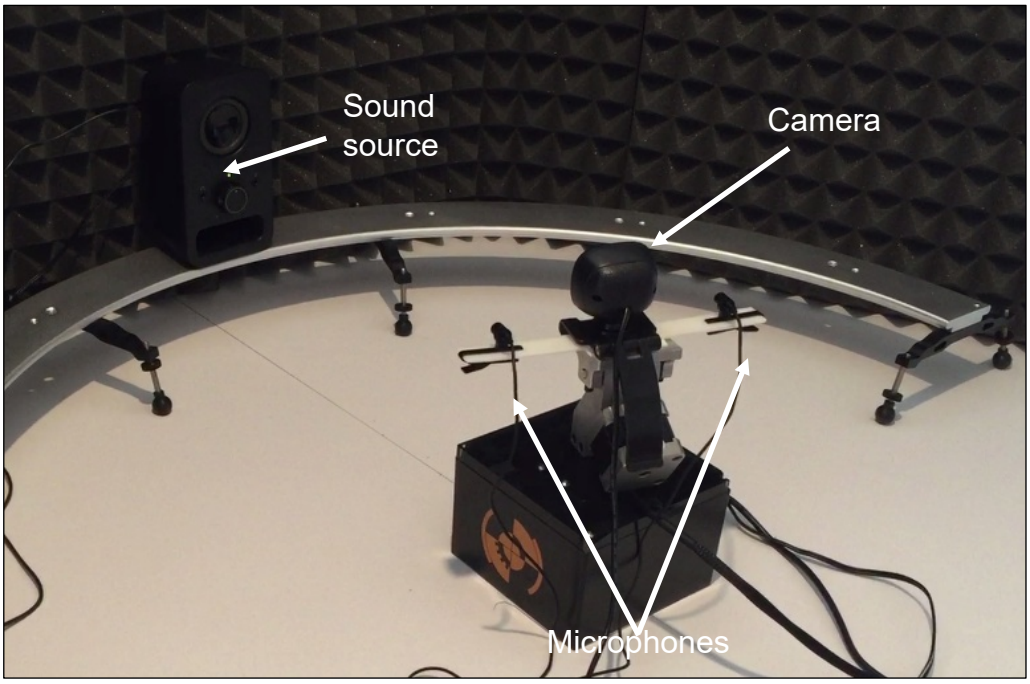


Figure 17. Experimental apparatus for the cerebellar calibration pilot experiments.

4.2.3 Performance measurement

There are a number of ways to measure performance of the proposed system, which fall into two broad approaches. First is to focus on accuracy of performance such as MSE (in this case, in SSL estimation). The other is to focus less on accuracy and more on the qualitative behaviour of the system, that is, whether a robot using this system behaves in a useful way. This thesis uses a mixture of both, with an emphasis on measures of accuracy, with a shift towards behaviour as the studies come to an end and there is a focus on further work, and the possibility of the system being used in a real world context. Two approaches have been used for accuracy measurement: MSE in the SSL estimation and accuracy rate. MSE is calculated as the mean squared difference between the estimated value (calibrated SSL estimate in the case of the calibration models) and the ground truth value

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_t - x_i)^2 \tag{10}$$

where x_t is the ground truth value (in the case of the SSL calibrators this will be an azimuth value), x_i is the i_{th} estimate and n is the number of trials (estimates). In Figure 20 and Figure 23, the Root Mean Squared (RMS) error is plotted, which is the square root of the MSE.

Accuracy rate is the percentage of SSL errors that fall within a certain range, and was inspired by Li et al. [34]. Li et al. used a criterion of percentage errors less than 15º, however, in this thesis, that threshold usually resulted in a 100% accuracy rate, making comparisons meaningless, and so a threshold of 5º error was used instead. Accuracy rates were still quite high, especially for multiple models with RP, which typically were around 98% to 100% and so it may also be worth investigating the use of lower threshold values, although there comes a point where the threshold value approaches the resolution of the SSL algorithm. Accuracy rate is calculated by counting the absolute error values (absolute value of the difference between calibrated estimate and ground truth) that fall below the threshold value and calculating this as a percentage of the total number of trials (estimates). On reflection, the usefulness of this measure is questionable, as it simply categorises errors as either below a threshold or not; it may mask nuances in the performance. It may have been more straightforward and clearer to only use the MSE which appears to be standard practice in the literature.

A more qualitative approach has also been used, for example, visually inspecting plots of responsibility signals (covered in Chapter 6-Chapter 10), to observe whether the system behaves as expected.

In order to confirm that models developed in the thesis perform significantly better than other techniques with which they have been compared, some post- model learning experiments were conducted and Student's t-test applied [116]. The t-test (specifically, here, the paired-sample t-test) is a hypothesis test that compares the means of two sets of trials or experiments conducted on a pair-wise basis under the same conditions. For example, in section 4.3.2, the mean of the errors in SLL estimation produced by the single cerebellar calibration model across a set of azimuths (here, the azimuths are uniformly distributed) is compared to the mean of the errors produced by the un-calibrated SSL algorithm, in a pairwise manner, across the same set of azimuths. The null hypothesis is tested, that the two sets of errors come from the same population- that is, there is no significant difference between the performances of the two techniques being compared. If the null hypothesis is rejected, we can be confident, to a degree (typically to a 95% confidence level), that the systems developed are producing significantly better performance than those techniques with which they being compared (e.g. cerebellar calibration versus un-calibrated SSL). The matlab ttest() function was used, which defaults to a 95% confidence level. This produces an h- value which is either 0 or 1. A

value of 0 indicates that the null hypothesis is not rejected, and a value of 1 that it is. The test also produces a t-value whose value indicates the size of the difference between the means relative to the variation in the differences between the means. A value close to zero indicates that the case to reject the null hypothesis is weak. The sign of the t-value indicates which of the samples has the smaller error; in this thesis, a negative t-value indicates that it is the proposed approach that has the smaller error than that with which it is being compared.

4.3 Results

4.3.1 Error introduced by a physical object

Results of one experimental run are shown in Figure 19 in which one microphone is obscured by an object (a box file), and it assumed that this will disrupt the time of arrival of sound at the right microphone (the experimental arena is shown in Figure 18). This causes a leftward bias in the estimated azimuth of the source (shown in red in Figure 19). This particular experiment was run over 50 iterations with the cerebellar weights (parallel fibre/Purkinje cell synapse weights) updated on each iteration. By inspection the calibrated position estimates (shown in green) appear broadly similar to the ground truth (shown in blue), although it is difficult to match up individual data points from this plot. It is worth noting that this data represents performance during learning, so that in this and other similar plots, there will be green data points from early in the learning that show poor performance, and others from later in the learning cycle that show better performance; also, the radial offsetting of the red and green data points (for the purposes of clarity) does make the data appear rather better than it actually is. Learning is successful within a few iterations, as shown in Figure 20. Figure 21 shows the results of using the learned weights post-learning with a fresh set of positions generated on a linearly spaced grid. This also demonstrates some generalisation as the leftmost ground truth point in Figure 21 represents data that was not used in training yet there is clearly appropriate correction.

4.3.2 Error introduced with artificial distortion of the SSL output

In order to quantify the performance of the cerebellar calibration compared to that of the un-calibrated SSL, a new experiment was run using artificial distortion as described in Section 4.2.1. Note that this experiment was actually run at quite a late stage in the project, and the recorded audio from Chapter 7 was used, as the setup described in Section 4.3.1 was no longer readily available. In this experiment, 60 iterations were used to train the

model then 50 post-training trials were performed with randomly selected sound source azimuth (each SSL output artificially distorted using the same algorithm as during training), to compute the MSE in the SSL estimate with and without calibration. The MSE of the artificially distorted, un-calibrated SSL was 14.8 degrees^2 and that of the calibrated output of the cerebellar model was 1.14 degrees^2, suggesting that the SSL performance with cerebellar calibration is markedly superior to that without. As described in section 4.2.3, the experiment was repeated with 81 uniformly selected azimuth values from, and a paired-sample t-test carried out on the calibrated and uncalibrated results, using the Matlab `ttest()` function. The h value was 1, suggesting that the null hypothesis (that the difference between the means of the calibrated and uncalibrated samples is zero) is rejected, and the improved performance of the calibration model can be treated as significantly better than that of the uncalibrated SSL estimate at the 95% confidence level. The standard deviation of the difference in mean errors was 1.9^o. The t-value was -13.4: the negative sign indicates that the calibrated errors were smaller than the uncalibrated errors and the value is not close to zero (the closer to zero, the weaker the case to reject the null hypothesis).

4.3.3 Performance with visual feedback

Figure 22 shows the results of training the model using sensory feedback (vision) to derive the ground truth, as discussed in Section 4.2.2, using artificial distortion as described in Section 4.2.1, with learning taking place over 50 iterations. The positioning of the green data points indicate that learning has taken place and this is confirmed by the decrease in error shown in Figure 23. Figure 24 shows an image taken from the camera during the experiment. The underlying assumption in using vision as providing the ground truth is that it is perfect, which of course, in the real world it is not. It may be realistic assumption however, that investigating this further as a sensor fusion problem, may yield improvements in successful robot behaviour.

4.3.4 Performance with pure tone

An experiment was carried out using a 500Hz pure tone (sine wave) for the audio source rather than the Gaussian noise used throughout the remainder of the thesis, to check that the system will deal with other types of sound. Artificial distortion was introduced as described in Section 4.3.2, and calibration models learnt over 60 iterations. Performing SSL with pure tones is difficult, especially for the simple SSL algorithm used in the thesis (indeed, many of the azimuth data points resulted in an invalid ITD value meaning that

azimuths were restricted to the range of -13° to 30° for this experiment); nonetheless, the performance of the calibrated system is impressive with an MSE of 6.6 degrees^2 compared to 200.5 degrees^2 for the un-calibrated SSL output.

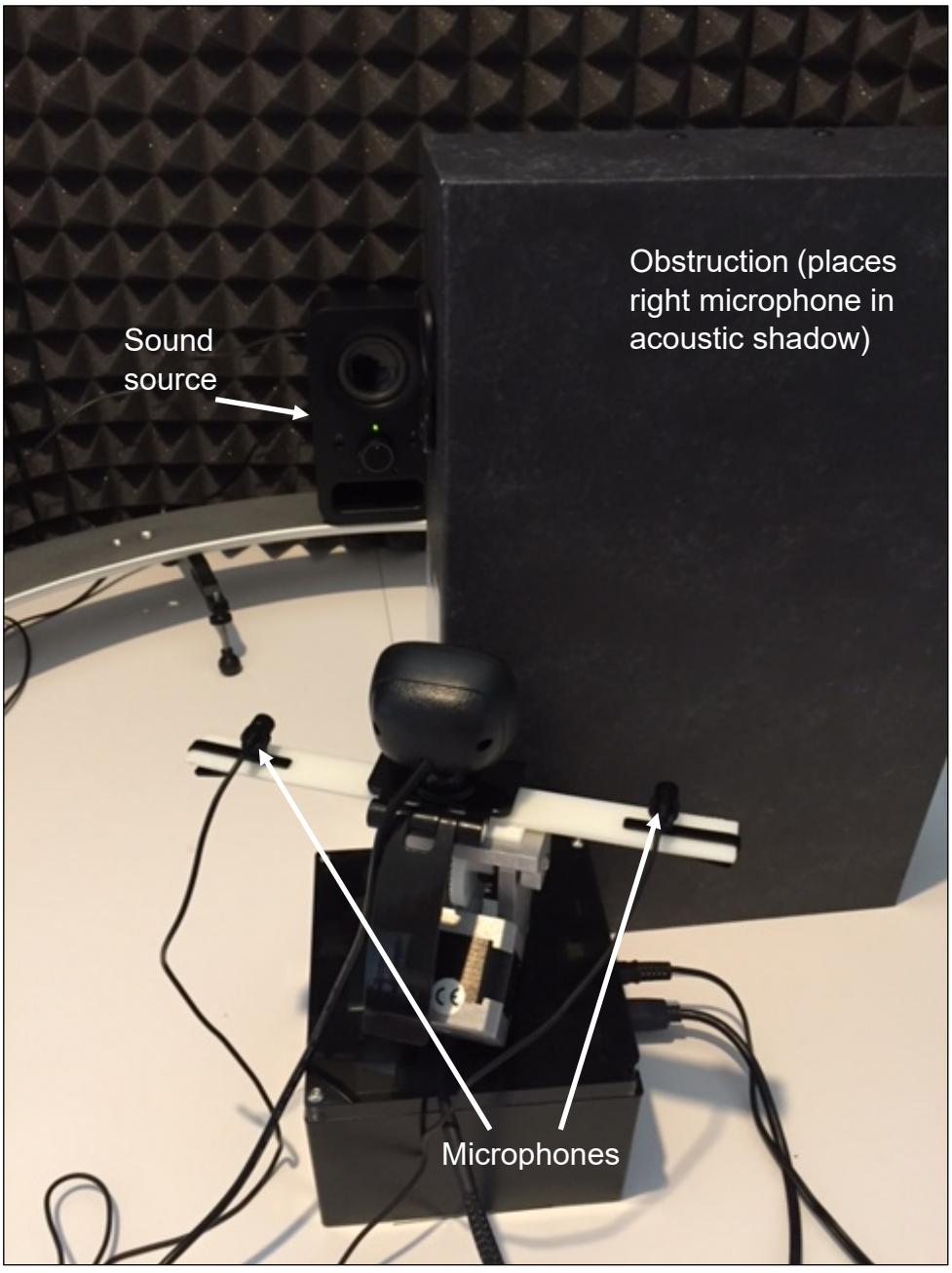


Figure 18. Experimental arena with obstruction.

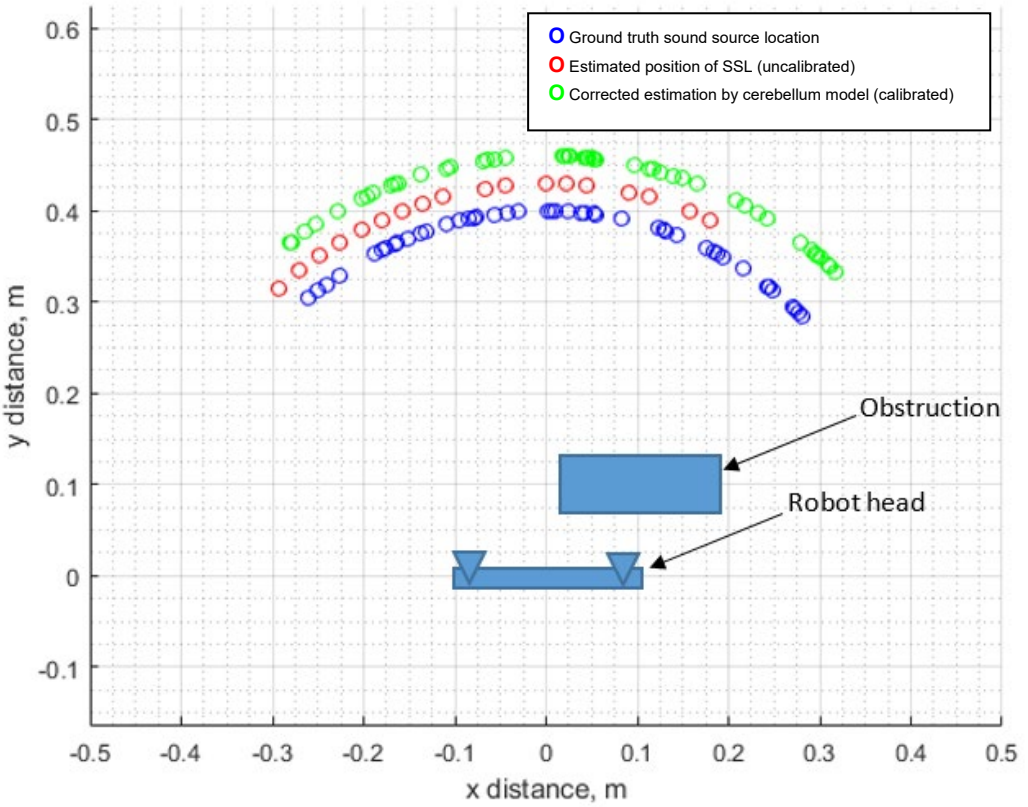


Figure 19. Results of cerebellar calibration in learning mode. One microphone is obscured. Estimated position (red) and calibrated (green) are offset in the y direction for clarity. Calibrated positions show learning (update of cerebellar parallel fibre/Purkinje cell weights) over 50 iterations.

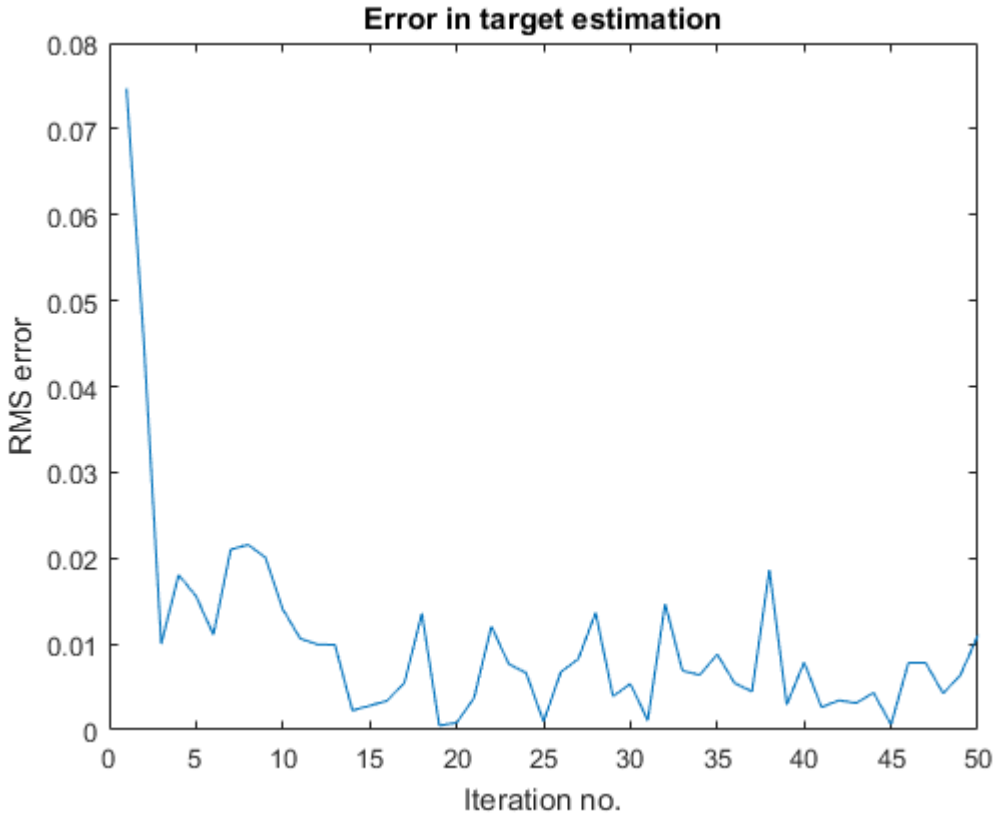


Figure 20. Error in target estimation during learning. y axis units are metres: distance along azimuthal arc from 0º.

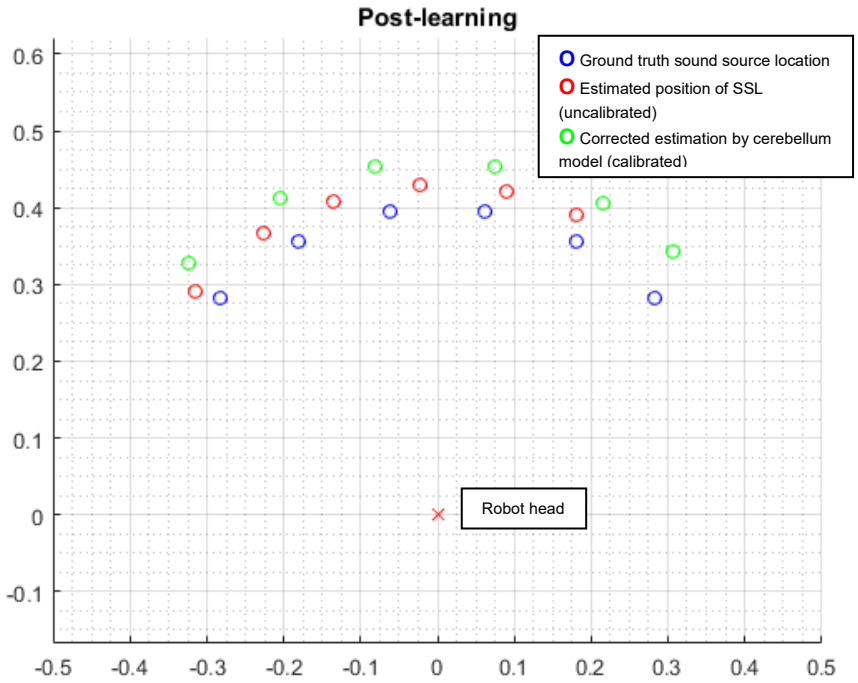


Figure 21. Results of cerebellar calibration post-learning. Axes show x,y distance in metres with robot head at the origin. Green points show calibrated estimates of source azimuth with learned weights applied to a new set of regularly spaced positions.

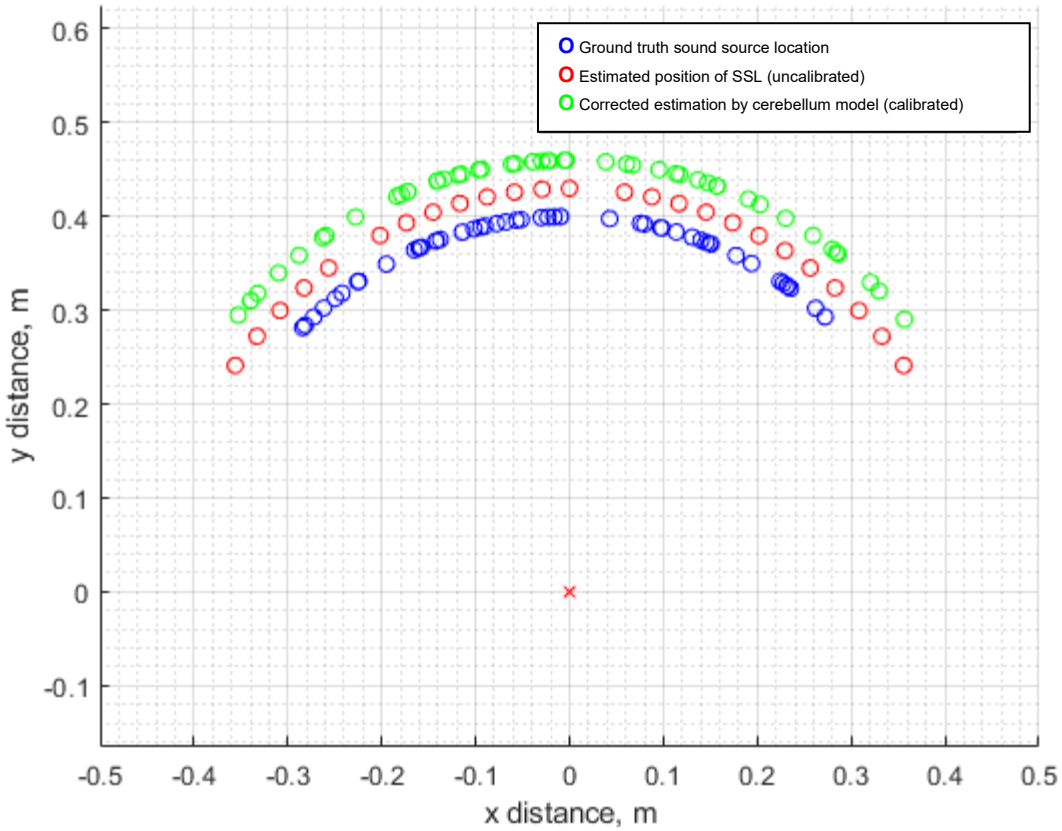


Figure 22. Results of cerebellar calibration in learning mode using vision. The SSL estimate is artificially distorted by multiplying SSL estimates by a factor of 1.3. Estimated position (red) and calibrated (green) are offset for clarity. Calibrated positions show learning (update of cerebellar parallel fibre/Purkinje cell weights) over 50 iterations.

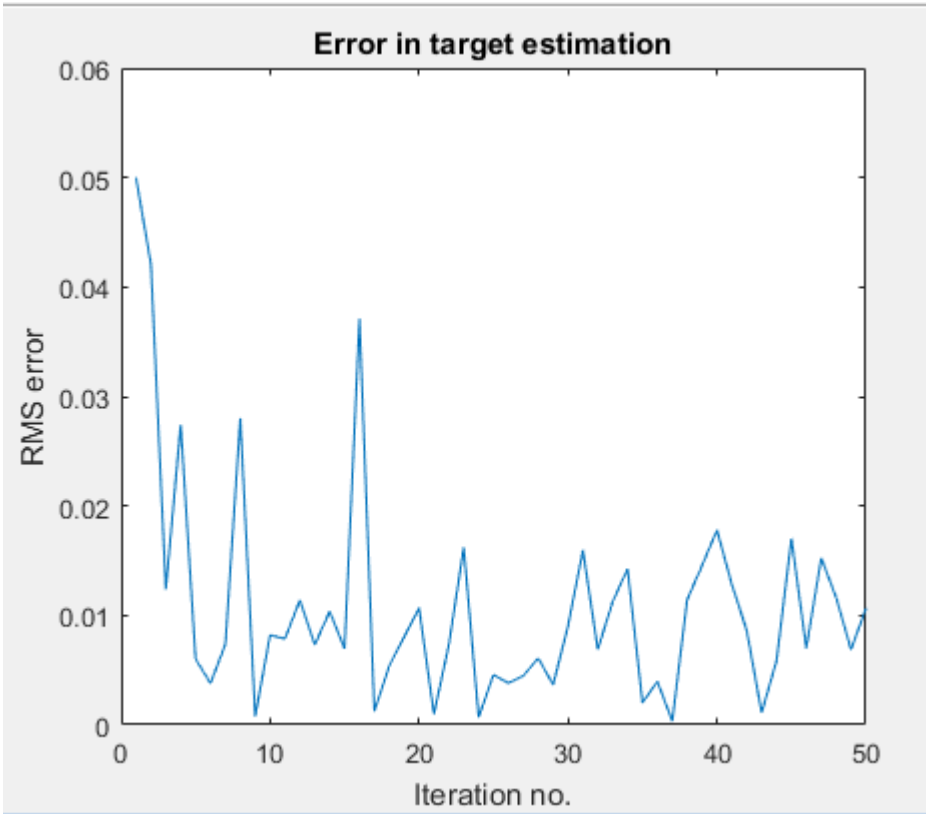


Figure 23. Error in target estimation during learning, using vision. y axis units are metres: distance along azimuthal arc from 0º.

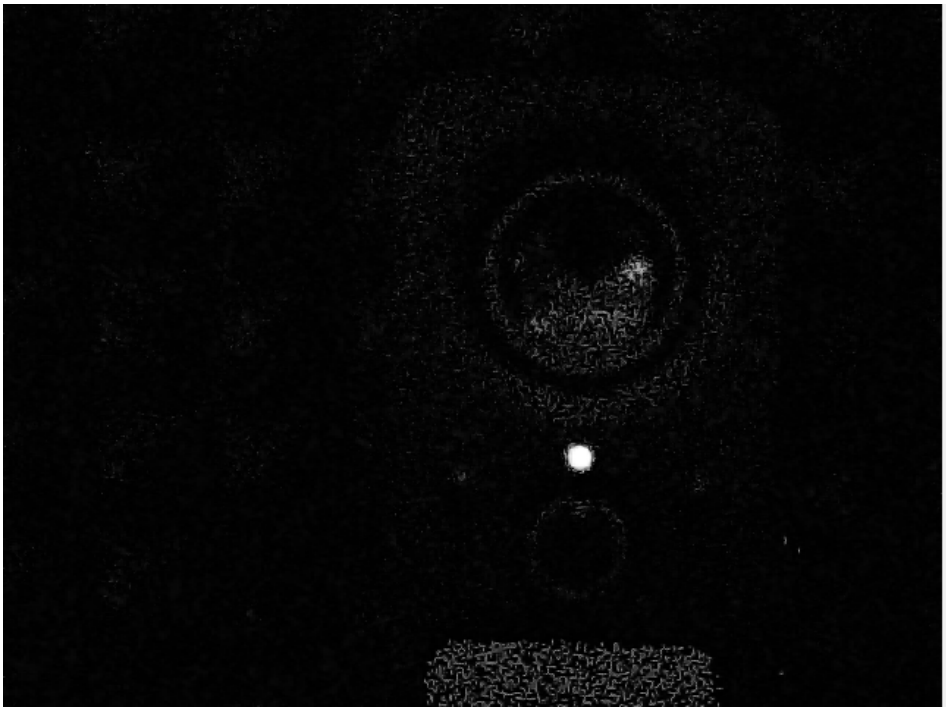


Figure 24. Image capture from camera during visual ground truth experiment.

4.4 Chapter summary

This chapter has demonstrated the basic SSL calibration technique based on the adaptive filter model of the cerebellum. The proposed system was able to learn to compensate for azimuth-dependent errors introduced into the SSL estimate due to imperfections introduced into the acoustic environment, and to do so with a relatively low number of learning trials (typically 20-60).

This was an important initial result as it forms the basis of the work in subsequent chapters, especially the multiple-models approach to SSL calibration.

The system also appeared to display generalization, successfully calibrating SSL estimates for sound source positions that were not included in training. The method of generating audio stimulus of various azimuths however is unrealistic (rotation of the robot head rather than displacement of the sound source itself) and so a means of automatically positioning the sound source itself, rather than mimicking this by rotating the robot head, is required. This is addressed initially in Chapter 6 with new experimental apparatus, and the apparatus is further redesigned for improved reliability and portability in Chapter 7.

Although Gaussian noise has been used as a sound source throughout the thesis, a limited experiment was performed to check that a different and more challenging type of sound, a pure tone, could be successfully calibrated, with remarkably successful results.

Finally, throughout the thesis the ground truth sound source azimuth is determined through the odometry of the experimental apparatus and it is proposed that a robot operating in the field could do so through the use of sensory feedback, in particular, vision. A pilot experiment showed that the calibration technique developed will successfully learn using visual feedback to generate a teaching signal.

Chapter 5 Multiple models and internal models

5.1 Internal models

Internal models simulate some aspect of a system in order to estimate or predict the response of that system to some stimulus or action. The internal model has come about in the context of motor control and forms part of a control system that generates some output, usually a motor command. There has been growing acceptance that the brain makes use of internal models for motor control and that they are likely to be located in the cerebellar cortex [74, 105, 117-121]. Models of the cerebellum have increasingly been used as adaptive controllers [104], and the role of the cerebellum in motor control is discussed in Section 3.6. In the context of motor control, internal models contain a paired forward and an inverse model. Input to the forward model is the motor command but could also include sensory input [74] and the output is a prediction of the consequences of the action given the current state and motor command as input as the forward model learns the dynamics of the controlled plant [6, 105]. In this way, the forward model overcomes the problem of large delays in sensory feedback. On the other hand, an inverse model of the system acts as a controller, providing feedforward control [104] as it produces the required motor command for a given desired state. Another way of viewing this is that the inverse model acting in series with the controlled plant together form an identity such that the output is identical to the input provided the inverse model is accurate. Forward models can be learned through experience, adjusting for example synaptic weights of a NN in response to an error signal obtained from the difference between desired state and actual state ascertained through sensory feedback. Inverse (controller) models are more problematic to learn, as the teaching signal (the desired motor command) is not usually directly available, otherwise there would be no need for an inverse model. Solutions to this problem are discussed in Section 3.6.

5.2 Multiple models

5.2.1 Introduction

As mentioned in Section 1.1 and 5.1 it is increasingly becoming accepted that the brain possesses internal models of the external world. These models allow the prediction of the way in which the world will behave, such as predicting the consequences of an action. A single model would not be able to capture the range of contexts encountered in real world

situations [122] and there have been a number of proposals that the animal's central nervous system makes use of multiple modules (containing models), each specialised for a specific context [6, 122-130]. That is, a module will contain models that have learned to make predictions about the external world and then take control of some activity in a specific situation, or context. Three key advantages are claimed for modularity [6]; first is the modular nature of the world that an animal (or robot) inhabits with objects and environments "parcelled up" into discrete contexts. Second, modularisation allows different modules to participate in learning without affecting the behaviours already learned by other modules, allowing new learning while retaining existing behaviours with little or no effect on those behaviours. Third, rather than learning a new behaviour afresh for each new context, the animal (or robot) can, for many such contexts, combine existing learned behaviours to cope with the new one. This could be achieved by a system that combines the outputs of modules in proportion to how well those models are suited to the new context.

There have been a number of such modular systems proposed, including *Modular Selection and Identification for Control* (MOSAIC) [6] and *Hierarchical attentive multiple models for execution and recognition of actions* (HAMMER) [131] and these were preceded by the *mixtures of experts* architecture [123]. These frameworks were developed in the context of motor control and action imitation. MOSAIC and HAMMER are similar to each other in that they are based on forward/inverse model pairs, and that the suitability of each to control in a particular context is determined through comparison of a state prediction with the actual state. A key difference is MOSAIC's RP that predicts its modules' suitability based on contextual signals, and this is lacking in HAMMER. Narendra et al. proposed a multiple model architecture in which the models were selected individually rather than having their outputs combined, that is, the model outputs were switched [126].

The MOSAIC framework was chosen in this thesis as it has a number of advantages including proportional combination of module outputs along with both prior and posterior contribution to the production of responsibility signals.

In this framework, a module consists of a forward model, which receives the efference copy of the motor command and makes a prediction about the consequences of it, concurrently with the forward models in the other modules. A separate *Responsibility*

Estimator (RE) operates across all modules and transforms the array of predictions into a set of signals, called *responsibilities*, one for each module, which represents the likelihood that each module is appropriate for the context and should be responsible for control. Each module in MOSAIC also contains an inverse model, which learns to produce an appropriate motor command in the context for that module, which forms the main output from the module. The other output from the module is the likelihood that the module is responsible for control in a particular context, which is computed using a prediction error generated using sensory feedback, and is input to the RE. The motor output from a module is modulated by its responsibility signal, provided by the RE and then summed with the other module outputs. The result is an overall motor command output from the system which is a combination of module outputs in proportion to each module's ability to control in the current context.

Frameworks such as MOSAIC were developed in the context of motor control and all examples in the literature reflect this. This makes the application of such frameworks to audio localisation somewhat problematic, and hence the thesis takes a multiple-models *inspired* approach rather than a faithful reproduction of such a system. Nevertheless, this chapter describes the MOSAIC system in its original context of motor control, and the adaptation of the approach to audio localisation is covered in later chapters.

The system needs to select the module appropriate to the context by switching the outputs of inverse models on or off (or modulating rather than switching if appropriate). This switching involves two processes [6]: first, is the generation of motor commands through the selection of the most appropriate controller (inverse model) for the estimated context based on sensory input. A second switching process uses sensory feedback of the consequences of the action to select a more appropriate model if necessary.

In MOSAIC, the inverse models' contribution is determined through a responsibility signal. This is derived through two further processes [6]: first, each forward model's prediction of the next state of the controlled system can be compared to the actual state through sensory feedback, but only after the action has taken place (or during action). The second process estimates responsibility from sensory contextual information, providing the potential to select modules before action.

5.2.2 MOSAIC system

The MOSAIC system was developed in the context of motor control and is introduced in [6]. The application of the systems in the thesis is in the context of cerebellar calibration of a robot's SSL algorithm, to which there is not a direct mapping- this will be covered in the discussion and for now MOSAIC will be described in its original context of motor control. There have since been a number of extensions including Hidden-Markov MOSAIC [122], Hierarchical MOSAIC [132] and Multiple Reward MOSAIC [133].

MOSAIC consists of an array of modules each of which could have influence or control in a particular context and is shown in Figure 25. Each module consists of a forward model, inverse model and a responsibility predictor. There is a single responsibility estimator that operates across all modules.

5.2.2.1 Input/output

The inputs to this system are (Figure 25):

- *Efference Copy*
- *Desired Trajectory*
- *Sensory Feedback*
- *Feedback Motor Command*
- *Contextual Signal*

Efference copy is a copy of the motor command currently sent to the motor system. This will be used in part (by the forward models within the modules) to predict the consequences of that command after action has taken place.

Desired Trajectory is the desired next state of the system and is used by each module to generate a motor command that would be appropriate for its context.

Sensory feedback provides the ground truth current state of the system under control. It is a means by which the MOSAIC system can ascertain the consequences of the previous motor command. This is used to determine how well each module was suited to provide that motor command, by comparison with the module's prediction of the current state. It is also used along with the *efference copy* by each module to make a prediction of the next state of the system under control.

Feedback Motor Command is the teaching signal for the component within each module (the inverse model- see Section 5.2.2.4) that generates the motor command for its context.

As mentioned in Section 5.1 the desired motor command is unavailable and some other means is required to generate a suitable teaching signal. This is not directly relevant to the approach taken in this thesis.

Contextual Signal conveys information about the environment. It is used to make a prior prediction about the suitability of each module to control in the current context. The signal is used to identify the current context before sensory feedback becomes available.

There is a single output from the MOSAIC system which is an overall motor command.

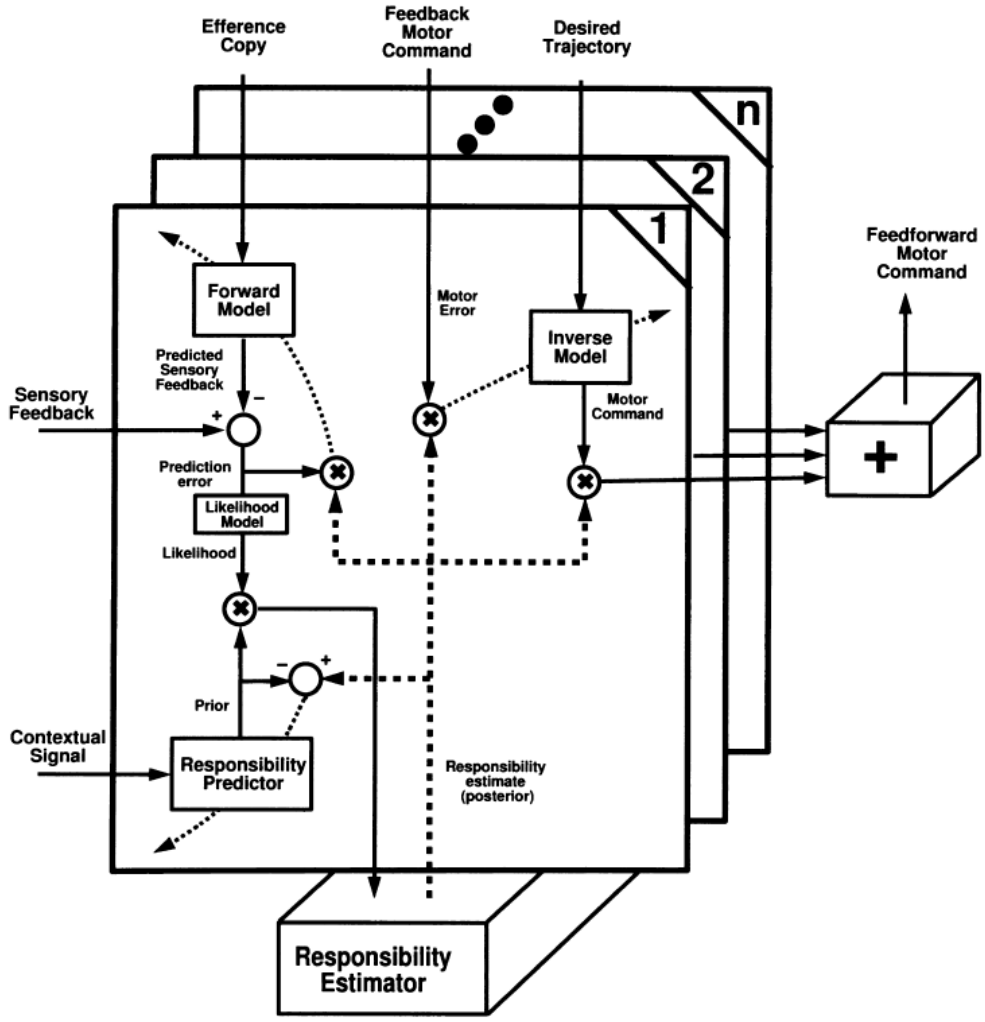


Figure 25. MOSAIC framework. Reprinted from [6] © 1998, with permission from Elsevier.

5.2.2.2 Forward model

There is no direct correspondence between the forward model (or indeed the inverse model, covered in Section 5.2.2.4), however, it is covered here for completeness, as it is part of the MOSAIC framework. There is an indirect analogy, in that the role of the

forward model is to predict the consequences of a motor command, assuming it is operating in a particular context, and the calibration model in this thesis is used predict the location of a sound source, assuming it is operating in a particular context. The forward model receives an efference copy- a copy of the efferent, or out-bound- motor command (the *Efference Copy* in Figure 25), and learns to predict the consequences of that motor command (i.e. it generates a prediction of the next state of the system). If the forward model is based on some sort of NN, then this learning would be through the updating of its synaptic weights through some learning rule, which would use the error in prediction. This error would be determined through sensory feedback of the actual consequences of action (the *Sensory Feedback* in Figure 25), and is transformed into the likelihood that the module is *responsible* for control in that context. This aspect of the module is shown in Figure 26.

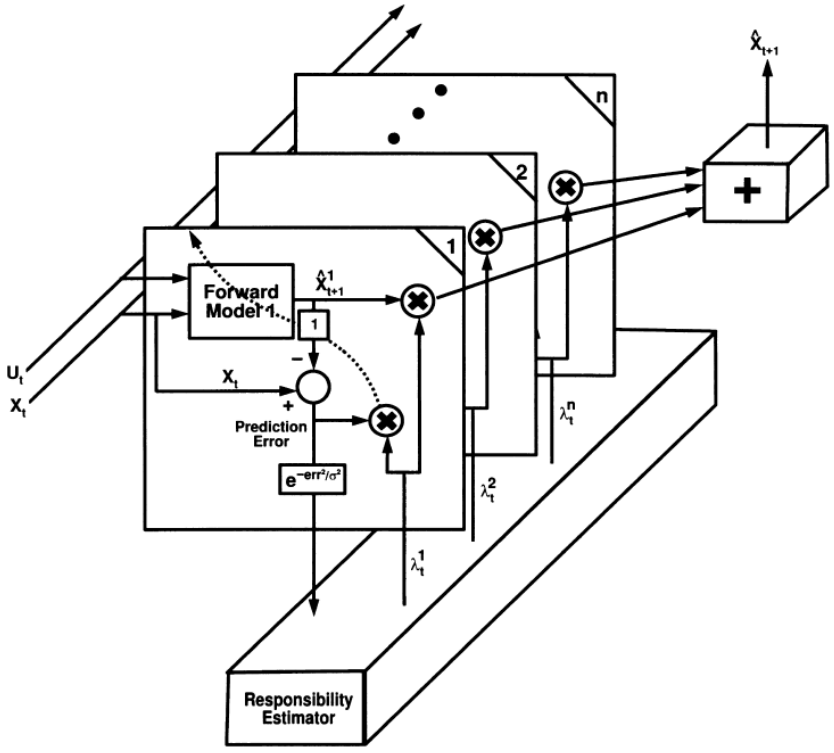


Figure 26. Multiple models showing only forward models. Reprinted from [6] © 1998, with permission from Elsevier.

5.2.2.3 Responsibility estimator

The *Responsibility Estimator* (RE) is a single unit that takes as input the likelihood values of all the modules. The set of likelihoods are used to make a decision about which module (if the outputs of modules are to be switched) or set of modules (if the outputs of modules

are to be combined) is best suited to control in a particular context by identifying the forward model(s) with the highest likelihood(s). The computation of likelihoods relies on sensory feedback about the true state of the system under control as mentioned in Section 5.2.2.2 and Haruno et al. refer to this as feedback selection of modules [122]. The likelihoods are derived from the prediction errors of the models, based on sensory feedback of the ground truth response of the system.

The likelihoods are normalised across all modules, using a *softmax* function, producing a responsibility λ_i for the i_{th} module:

$$\lambda_i = \frac{e^{-|x_t - x_i|^2/\sigma^2}}{\sum_{j=1}^{n} e^{-|x_t - x_j|^2/\sigma^2}} \tag{11}$$

where x_t is the true state, x_i is the estimate produced by the i_{th} model, n is the number of models and σ is a scaling factor which is equivalent to the standard deviation assuming a Gaussian distribution of predictions. The responsibility signal λ_i has a value between 0 and 1 for each module which sum to unity across all modules. The value of σ determines the distribution of responsibilities across modules. Large values of σ will cause more even sharing of the responsibilities across modules, whilst a smaller value will accentuate those modules with the highest likelihoods causing them to dominate control. It might be tempting to use a small value of σ to switch to the module with the highest likelihood and have that module dominate control, but this would diminish the system's ability to generalize to novel contexts, so a trade-off needs to be found between too low a value of σ at the expense of generalisation, and too high a value that might result in near-equal sharing of responsibilities (which would then be no better than having a single module).

The responsibility values are used in two ways. First, across the modules, the responsibility values are used to sum module outputs (motor commands) in proportion to the responsibility values to produce an overall output (motor command)- covered in Section 5.2.2.4. In the context of this thesis, this will be a calibration signal. This combination of modules (rather than simply the identification of the best module) in proportion to their ability to control allows the system to cope with novel contexts whose features that lie intermediate to those of the contexts in which each model has been trained. As such, modules should be able to be combined such that the system can deal with more contexts than there are modules [6]. However, MOSAIC can only interpolate

to novel contexts, that is, cope with contexts whose characteristics lie intermediate to familiar contexts. The system cannot extrapolate to contexts whose characteristics lie outside the learned state space. Second, the teaching signal for the forward model is modulated by its module's responsibility signal, so that those models with higher responsibility values receive more of the teaching signal. This is significant for two reasons. First, it allows a de-novo system to divide up experience through the competitive learning of the modules. Second, if the system encounters a context that is only slightly different from one encountered before (rather than a completely novel context), it allows "tuning" of the models best suited to that context (whilst others remain unchanged). This does leave a question, unanswered in the literature covering MOSAIC, as to when a module should adapt to a novel context (that is similar to a familiar one) and when a new module should be produced to learn afresh in the new context.

As the prediction error cannot be found until sensory feedback becomes available, the responsibility computed using (11) is posterior. This is fine as long as the context remains the same or perhaps changes only slowly. However, if the context changes abruptly, this could result in a relatively high performance error, as the "wrong" responsibilities will be used until they are updated through sensory feedback, at which point the system will adjust the blend of responsibilities to mitigate the error.

5.2.2.4 *Inverse model*

As with the forward model, there is no direct correspondence between the calibration models developed in this thesis and the inverse model in MOSAIC, but again, there is an indirect analogy in that the inverse model learns to control in a particular context, whilst the calibration model developed in this thesis learns to calibrate in a particular context. This aspect of the module is shown in Figure 27.

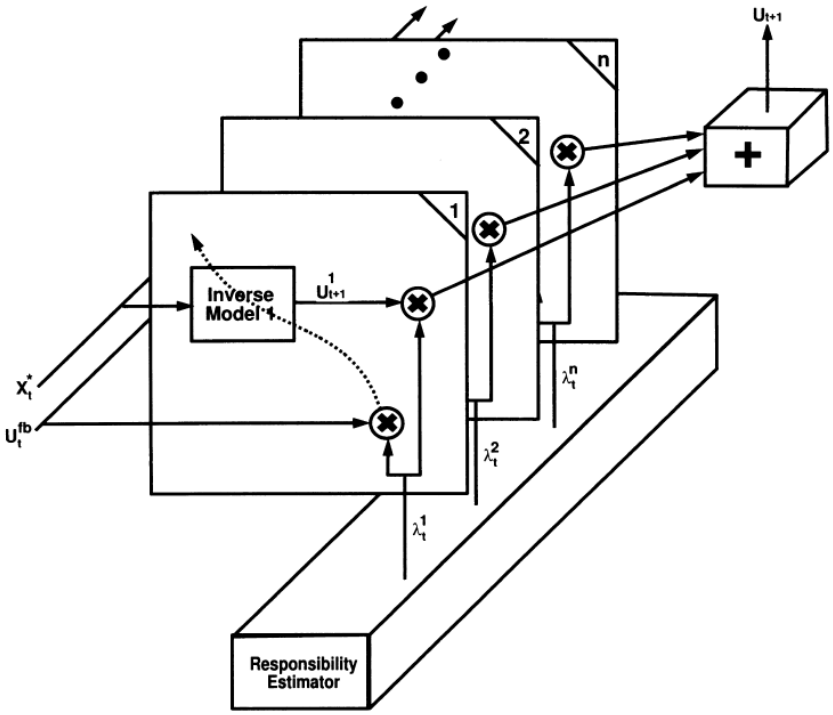


Figure 27. Multiple models showing only inverse models. Reprinted from [6] © 1998, with permission from Elsevier.

The input to the inverse model is the desired next state of the system and it learns to produce the motor command to achieve that state through comparison of its motor command with the desired motor command. This, of course, is biologically implausible as the desired motor command would be unavailable, and if it were, there would be no need for the inverse model. Wolpert et. al suggest that feedback-error-learning, in which a feedback controller (a linear approximation of the inverse model) uses negative feedback to produce a motor error [134]. As described in Section 5.2.2.3, the inverse models' outputs are summed in proportion to the modules' responsibility values to form an overall motor output. As with the forward model, the teaching signal for the inverse model is modulated by the responsibility signal for its module.

5.2.2.5 Responsibility predictor

As mentioned in Section 5.2.2.3 the responsibility values computed by the RE are posterior, that is, they are based on sensory feedback of the consequences of action. This causes a potential problem, mentioned in the same section, that rapid changes in context will not be detected until sensory feedback becomes available. This means that the system would still be using the responsibility values from the previous context so that there could be a large performance error as the system fails to adapt quickly to the new context (a

potentially catastrophic situation). Wolpert and Kawato propose a *responsibility predictor* (RP) in [6], which learns to predict its module's responsibility value based on contextual cues, producing a prior probability of responsibility λ_{pi} for the *i*th module. This is shown in Figure 28.

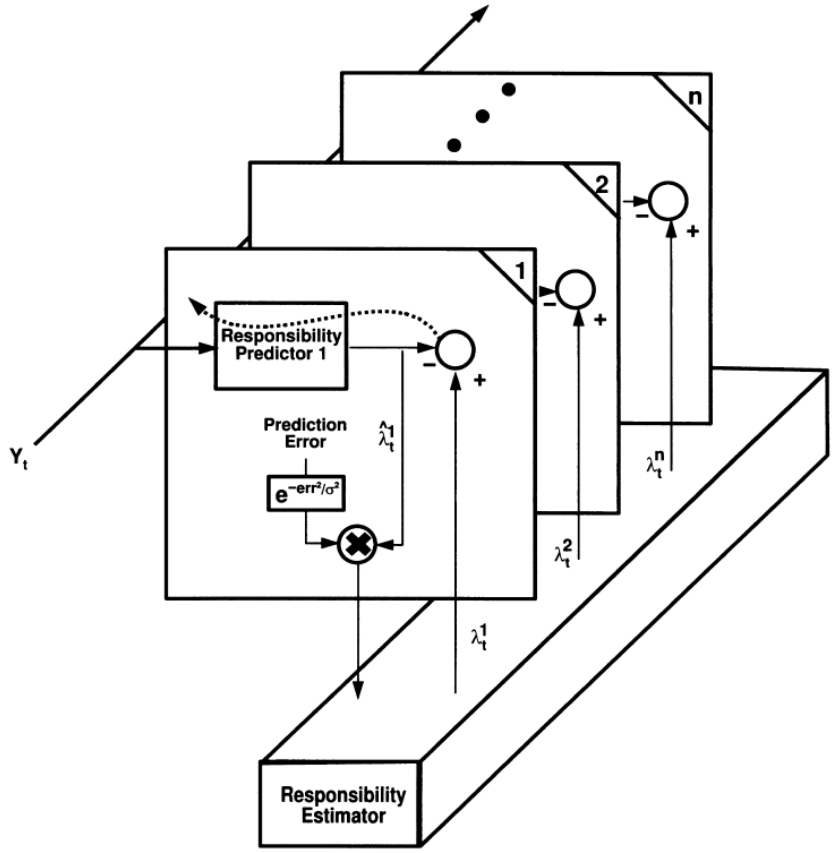


Figure 28. Multiple models showing only responsibility predictors. Reprinted from [6] © 1998, with permission from Elsevier.

The RP output is combined with the posterior responsibility, to give an overall responsibility:

$$\lambda_i = \frac{\lambda_{pi} e^{-|x_t - x_i|^2/\sigma^2}}{\sum_{j=1}^{n} \lambda_{pj} e^{-|x_t - x_j|^2/\sigma^2}} \tag{12}$$

This enables the system to correct itself *a-posteriori* if the error due to the RP is large. Conversely, if there is a large error in posterior prediction, perhaps due to an abrupt change in context as discussed in previous sections, this is mitigated by the prior responsibility. Indeed as discussed in Section 10.4, the sensory consequences of action could in some situations be unavailable (although this is not covered in the MOSAIC

literature), making the computation of posteriors impossible, so that the system can fall back on the RP which uses contextual signals that may still be available.

The approach is akin to a Bayes filter and a somewhat passing reference is made to this in [122]. The responsibility value would be analogous to the posterior probability in a Bayes filter, the RP output would be the prior and the RE output would be the evidence.

Contextual signals could be of any nature that allows a prior prediction of the responsibility of each model and in the context of motor control would typically be visual. For example, if we are about to lift an object, we can visually identify it as being light or heavy, preselecting appropriate modules. In Chapter 8, the audio stream itself is used to derive contextual signals, in the form of audio features.

5.2.3 Hidden-Markov MOSAIC

Haruno et al. introduce a Hidden-Markov Model (HMM) in [122], so that unequal state (context) transition probabilities are taken into account (in the article, it is implicit that all state transitions are equally probable) and it is claimed that this improves switching of modules, especially for low-frequency switching of contexts. HMMs were developed in the context of speech recognition and the technique is summarised in [135]. This version of MOSAIC takes the probability of a particular sequence of contexts into account, rather than treating contexts as isolated, unrelated states. Rather than compute the likelihood that a module is responsible for a particular context, the likelihood that the module is responsible for a particular *sequence* of contexts is computed as

$$L(X \mid \theta) = \sum_{\xi} \prod_{t=0}^{T-1} a_{s_{t-1}s_t} L_{s_t}\left(X(t) \mid w_{s_t},\, \sigma_{s_t}\right) \tag{13}$$

Where θ is the model parameters (e.g. NN weights), s_t represents a module selected at time t, ξ is all the possible sequences of responsible modules, L_{s_t} is the likelihood that the module selected at time t generates the sequence of states and $a_{s_{t-1}s_t}$ is the probability of transition from state s_t to s_{t-1}. A potential advantage of this approach is that it allows the automatic determination of the scaling factor, which is hand-tuned in the original MOSAIC, but in HMM-MOSAIC is determined through the Expectation-Maximisation (EM) algorithm. EM is a two-step iterative process. In the first step of an iteration, known as the E step, an initial guess is made of the parameters of the likelihood function and the observed data (in this case, the sequence of contexts) computed using this guess. In the

second step of the iteration, known as the M-step, new parameters are computed based on the estimates given the observed data. The Iteration continues until convergence. For practical reasons of computational tractability (due to the large number of possible sequences of contexts), the likelihood is reduced to

$$L(X \mid \theta) = \sum_{i=1}^{n} \alpha_t^H(i)\beta_t^H(i) \tag{14}$$

where n is the number of modules, $\alpha_t^H(i)$ is forward probability, the probability of the observed sequence to time t, and $\beta_t^H(i)$ the backward probability, which is the probability of the sequence observed from $t+1$ to the end of the sequence. The responsibility of a module is now

$$\gamma_t(i) = \frac{\alpha_t^H(i)\beta_t^H(i)}{\sum_{i=1}^{n} \alpha_t^H(i)\beta_t^H(i)} \tag{15}$$

The original MOSAIC was used as an inspiration in this thesis, as a "proof of concept", however, HMM-MOSAIC may be a fruitful area of investigation for future work, with the promise of automatic determination of the softmax scaling factor being particularly compelling.

5.3 Chapter summary

A weakness of the original MOSAIC is that the scaling factor σ in Equations (11) and (12) is tuned by hand, and the distribution of responsibilities is sensitive to its value. One could hand tune the value for a given set of contexts, but there is then the question as to how a system would operate autonomously. HMM-MOSAIC could be a way to automatically calculate this, and this may prove a fruitful area for future work.

A key issue not addressed in the MOSAIC literature is that the number of modules is chosen manually (based on the number of contexts), and so is fixed for a given set of contexts. Although the literature describes how the system can adapt to novel contexts (the best suited modules will receive more of the teaching signals and will adapt), it does not address how a decision is made as to when a new module needs to be created.

Despite the weaknesses identified in the basic MOSAIC described in [6], since the thesis represents a fundamentally different area of application, it would seem sensible to use this basic version of MOSAIC, recognising its limitations, to demonstrate the concepts of the thesis, and extensions and refinements to MOSAIC could inform future work.

Chapter 6 Acoustic context estimation using parallel cerebellar models

6.1 Introduction

This chapter represents the first experiment using multiple, parallel cerebellar models, and focuses on the use of the models to detect the acoustic environment (henceforth referred to as the *context*) in which the robot is operating. This work was published in the *Proceedings of the 18th Towards Autonomous Robotics* (TAROS) conference [113]. The main contribution of this chapter is the development of *responsibility estimation* as applied to the proposed system.

The work in this chapter benefits from a significant development of the experimental apparatus as a summer intern project at Bristol Robotics Laboratory. Whereas in Chapter 3 the sound source remained stationary, and the robot head rotated to mimic displacement of the sound source, here, apparatus was constructed (some of it as part of the intern project mentioned above) to allow automated positioning of the sound source in a more realistic way in relation to its environment, as explained in section 4.4. In addition, a means of rotating the sound source (loudspeaker) on its vertical axis was developed, in order to introduce SSL errors in a more reproducible way and under computer control. By orienting the sound source away from the robot head (Figure 31), the direct path to the robot head is less likely to be dominant, with indirect paths becoming more prominent, in a manner which should depend on the angle at which the sound source is oriented. For example, if the sound source is oriented at an angle ϕ of 45° in Figure 31, the sound wave front will be directed toward, and reflect off, the screen surrounding the arena. The smaller the value of ϕ, the more direct the path that will tend to be taken by sound waves to the robot head. The nature of the reflections will be unpredictable due to imperfections in the construction of the arena and also due to the front of the arena being open, so that, to an extent the arena acoustics will be affected by the room acoustics in which it is situated. This is in fact desirable, because in a real-world scenario, the error in SSL estimation is likely to be dependent on the sound source azimuth, as the path taken by the sound wave to the robot head would change depending on the azimuth.

The system is shown in Figure 29. This is a simplification of the multiple models framework, implementing only the models and the RE, which simply attempts to identify

the most appropriate model for a given context. The RE generates a set of outputs (1 output for each model), which represent the individual likelihoods that models are responsible for a particular context (i.e., that a model is the most suitable for that context). Identification of the context per se is not the focus of the overall system (rather, it is the improvement in the localisation performance using multiple models) but is the focus in this chapter to test whether the system is able to differentiate between the different contexts.

Each cerebellar model, having learned in a particular context, produces a calibration signal based on the output of the SSL algorithm (as explained in Chapter 4), according to Equation (8), which should depend not only on the azimuth of the sound source (assuming that the error introduced is azimuth-dependent, which it may or may not be but generally was found to be in this thesis), but also on the context within which the model has learned. Each calibration signal is then added to the estimated position produced by the SSL algorithm to produce a calibrated estimate of the sound source location- one estimate for each model. Hence, a set of azimuth estimates are produced from a single SSL estimate, and the idea is that the characteristics of the different environments in which the models learned will be reflected in the different estimates produced. The problem is then one of how to identify the correct context. It is assumed that the model that has learned in the current context will produce the lowest error in azimuth estimation (of course, this is not always the case, as discussed in Section 6.3. Each calibrated azimuth estimation is compared to the ground truth position, which is already known from the positioning of the sound source. Although, of course, in the real system, the ground truth cannot be found until the robot head orients toward the sound source, it has been used here merely for convenience to test the efficacy of the approach, and would ultimately be used with visual feedback on a mobile platform. The resulting prediction error is treated in a similar way to the MOSAIC errors and Equation (11) is adapted using azimuth rather than plant state in MOSAIC, so that the responsibility of each model is calculated as

$$\lambda_i = \frac{e^{-|\theta_t-\theta_i|^2/\sigma^2}}{\sum_{j=1}^{n} e^{-|\theta_t-\theta_j|^2/\sigma^2}} \tag{16}$$

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 equivalent to the standard deviation assuming a Gaussian distribution of estimates, and is set to unity in this specific

configuration. The maximum soft-max value (responsibility) determined through this computation corresponds to the lowest error in estimation and is assumed to correctly identify the context. The value of σ determines the distribution of responsibilities across models and has no effect on this identification, and so its value is not important in this particular study (however, it *will* be important where the outputs of models are to be combined in some way such as described in Chapter 7).

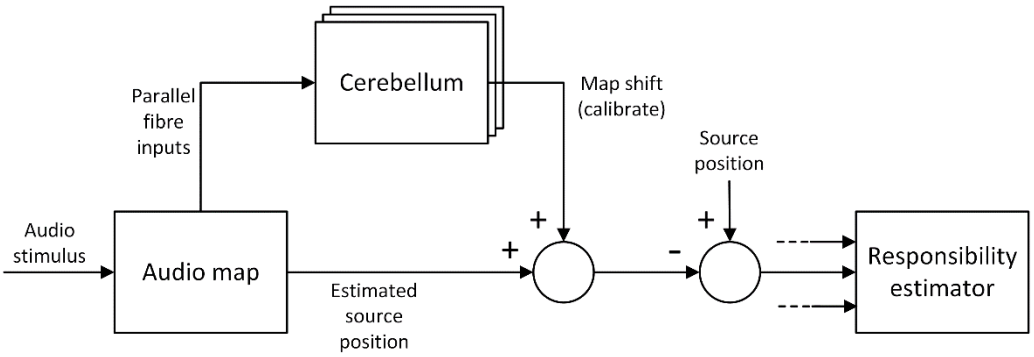


Figure 29. Multiple-models- inspired context estimation. For a given context, each model provides an estimate of source position. Each estimate is then compared to the ground truth source position and the RE classifies the acoustic context based on the estimation errors. Reprinted by permission from Springer Nature [113] © 2017.

6.2 Method

Algorithms were implemented in Matlab, and this was also used to control the running of experiments. The microphone setup was as described in Section 4.2, although with a larger inter-microphone distance of 0.25m. The inter-microphone distance was increased from that used in Chapter 4 to improve the SSL resolution.

The sound source was mounted on a motorised platform that could traverse a curved track such that it could be placed (under computer control) at any azimuth between -90º (left with respect to the robot head) and +90º (right with respect to the robot head) at a constant distance from the robot head (Figure 30). A curved photographic dolly system was used to constrain the path of the sound source to an arc, with the sound source itself mounted on a plywood platform fixed to the dolly mechanism. The rear of the platform engaged via a cog fixed to a stepper motor with a toothed belt which was fixed along the length of a semi-circular Perspex structure. A geared stepper motor was used in order to produce

enough torque to move the platform and this also allowed the source to be placed with a high resolution. 1° increments were used in this thesis although results are limited by the resolution of the ITD module, which varies from ±1.7° at zero azimuth to ±5° at ±70° azimuth. The motor was controlled via an auxiliary connection to the motor control circuitry in the PTU that was used in Chapter 4. The apparatus was quite difficult to use, with frequent adjustment of the relative positions of the two tracks and tensioning to arrive at a workable trade-off between sufficient tension to maintain satisfactory engagement between the motor's cog and the toothed belt, and the ability of the motor to drive the platform (too great a tension would result in the motor stalling).

The sound source was also mounted on a further stepper motor such that it could be rotated about its vertical axis through an angle ϕ as shown in Figure 31. This allowed the alteration of the acoustic context by rotation of the sound source so that it might face away at angle ϕ with respect to the robot head. This was a low current stepper motor that was controlled via an Adafruit motor controller board hosted on an Arduino board. This was controlled via USB from Matlab. The sound source could be placed at an orientation with respect to the robot head in 1.8° increments, the resolution of the stepper motor. The experimental arena was surrounded by a semi-circular screen that, combined with different orientations of the sound source, produced different acoustic contexts.

The cerebellar models were trained in different acoustic contexts. During learning, the robot head was presented, in each context (that is, for a constant value of source angle ϕ), with audio stimulus from randomised positions along the circular track, such that the direction of arrival of sound was from various azimuths (θ in Figure 31). 60 iterations were used to train a model. Post learning, all models were presented with the same set of audio stimuli at azimuths from -40° to +40° in 10° increments. For each stimulus, all models produce a calibration signal from which a set of errors are derived by computing the difference between each calibration signal (added to the SSL output) and the ground truth azimuth, and the soft-max of the likelihood for each model computed using Equation (16). Following the MOSAIC framework, the maximum soft-max, assumed to correspond to the minimum error, is used to identify the context.

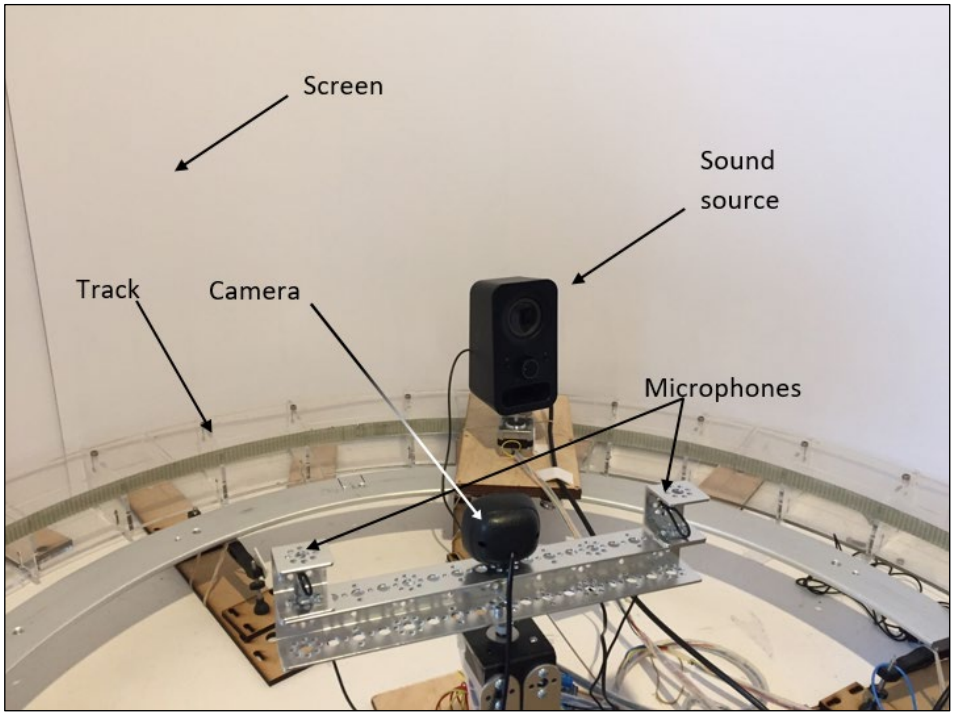


Figure 30. Experimental apparatus using track-mounted sound source. Adapted by permission from Springer Nature [113] © 2017.

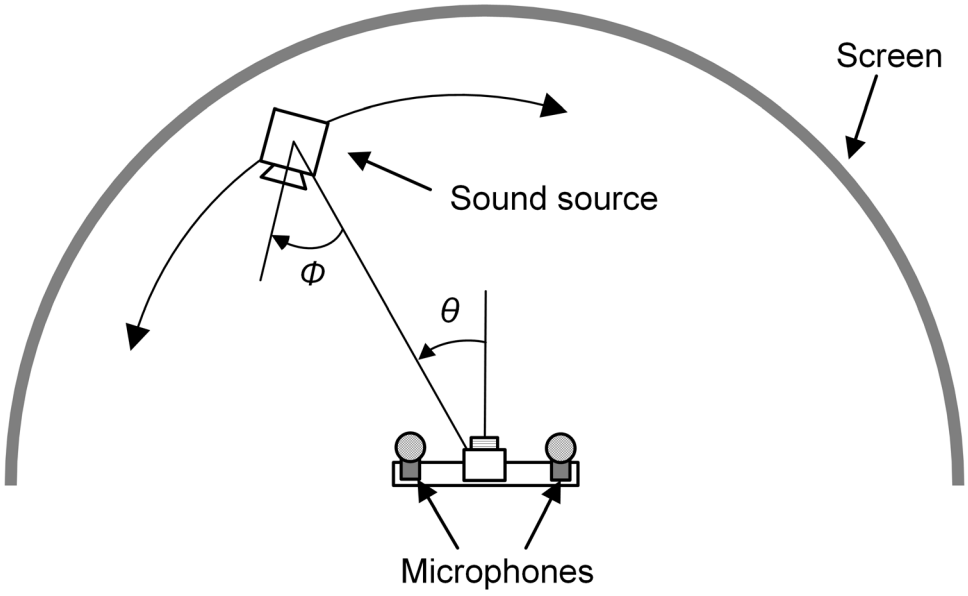


Figure 31. Plan view of the experimental apparatus. The source can be placed at various azimuths θ with respect to the robot head. The sound source can also be rotated at an angle ϕ on its axis. Reprinted from [54]. © 2018 IEEE.

6.3 Results

Seven cerebellar models were trained, as described in Section 6.2 with the sound source facing away from the robot head at a different angle (ϕ in Figure 31) for each model (135° left; 90° left; 45° left; 0°; 45° right, 90° right and 135° right with respect to the robot head). After training, the robot head was presented with sound source azimuths (θ in Figure 31) of 45° (left with respect to the robot head) to +45 ° (right with respect to the robot head) in 15° increments in each of the seven contexts. Therefore, an overall set of 49 (7 contexts, ϕ each with 7 azimuths, θ) different configurations where explored. For each source azimuth/context combination, the seven cerebellar models generated estimates of the context as described in Section 6.1, and the model with the lowest error was used to identify the context. Table 1 shows the rate of context identification. Each row in Table 1 represents seven different source azimuths in the same context.

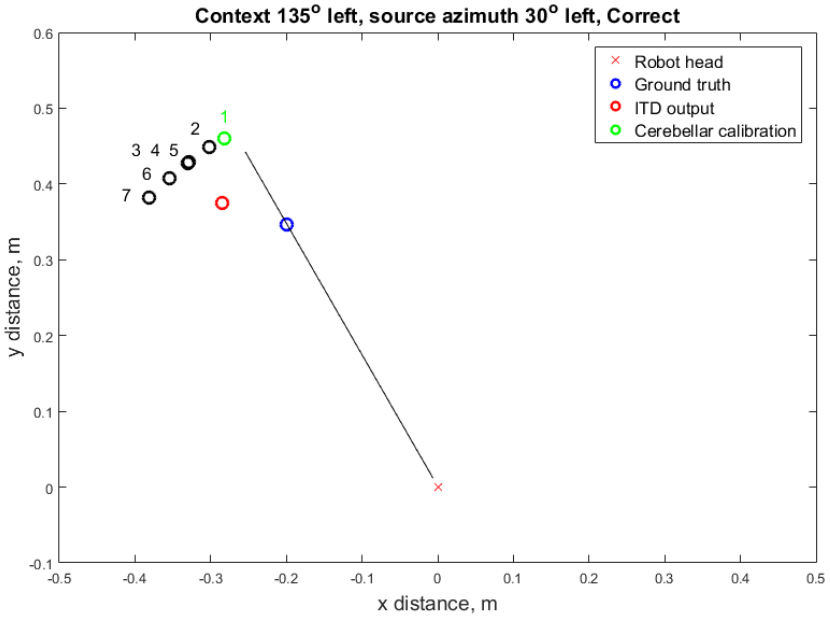
Table 1. Context identification with multiple models. Adapted by permission from Springer Nature [113] © 2017.

Context	Context (source orientation ϕ)	Correct identifications (n=7 azimuths θ)
1	135° left	85.7%
2	90° left	71.4%
3	45° left	42.9%
4	0° facing the robot	14.3%
5	45° right	71.4%
6	90° right	100.0%
7	135° right	100.0%

Figure 32 shows plots of sound source azimuths along with SSL estimates and cerebellar calibration by each of the seven models in one case in which context identification was correct (Figure 32a) and one case where context identification was incorrect (Figure 32b). The context is that the sound source is rotated (ϕ in Figure 28) 135° to the left away from the robot. The sound source azimuth (θ in Figure 28) is 30° left with respect to the robot head in Figure 32a and 30° right with respect to the robot head in Figure 32b. The context is the same for both plots, and so it should be the same model (model 1 in this case) that gives the lowest error. Blue circles represent the ground truth azimuth and red circles represent the un-calibrated SSL estimate. The green circle represents the estimate for the model that has learned in the given context.

The black circles represent the estimates of the remaining six models (those that had learned in the other contexts). It can be observed in Figure 32a that the estimate of the model that had trained in that context (model1) is the closest in value to the ground truth, leading to a correct identification, whilst in Figure 32b, a different model, model 2, is closer, so that the context was misidentified as that in which model 2 has learned.

(a)



(b)

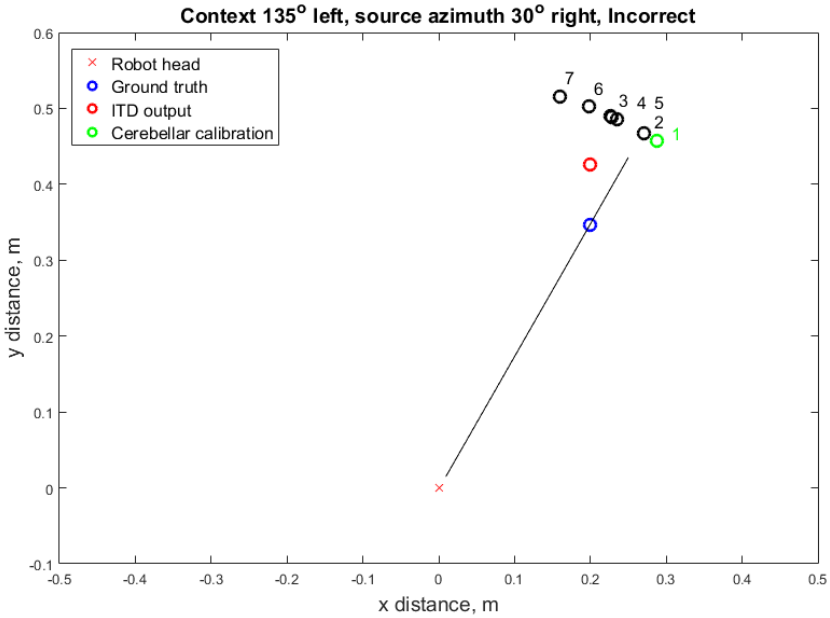


Figure 32. Plots of sound source azimuth. (a) Correct identification. (b) Incorrect identification. Adapted by permission from Springer Nature [113] © 2017.

6.4 Chapter summary

This chapter has presented a context estimation system which is able to identify the robot's acoustic context (albeit in a highly constrained way) with a degree of success, correctly identifying the acoustic context in 69.4% of 49 cases tested. Perhaps more importantly, however, the chapter demonstrates the basic utility of adopting the MOSAIC-inspired approach to SSL calibration, although that was not the focus of this chapter. Table 1 shows that the majority of contexts were correctly identified, and, where mis-classification occurred, this was mostly of a neighbouring (similar) context. The performance of the RE varies with the nature of the context. Mis-identification of the context more often occurs where there is little error in the SSL estimate and hence little difference between the model estimates. This is evident where the sound source directly faced the robot head (ϕ=0^{o}), so that all models produced similar estimates. The identification rate in this case was only 14.3%, no better than chance (the probability of selecting one of the 7 contexts, assuming that each has an equal chance of selection). Confusion can also occur where an "incorrect" model (that is, one that has not learned in the presented context) happens to produce a smaller error than the "correct" model (that is, the one that *has* learned in the presented context) as seen in Figure 32b. Success was greatest where the sound source faced away from the robot head, and there was a clearer distinction between contexts. In terms of localisation of the sound source, however, this may not matter, as the overall goal in the thesis is to identify the most appropriate model-even if that model did not learn in the presented context, and this is the subject of Chapter 7. Although the apparatus described in this chapter allowed the automated placement of the sound source, it proved to be unreliable, with frequent adjustment of the sound source locomotion mechanism and its tension being necessary during experiments. For subsequent chapters this apparatus was abandoned in favour of a redesign which is described in Chapter 7.

Chapter 7 Audio localisation using multiple models

7.1 Introduction

This chapter builds on the work developed in Chapter 6 to combine the outputs of the parallel cerebellar models. The work in this chapter was published in *IEEE Robotics and Automation Letters* [54].

The outputs of the models are combined in such a way as to improve calibration of the SSL in different acoustic contexts. This is a key chapter in that it demonstrates that the combination of the outputs of multiple adaptive filter models of the cerebellum, in the same way that the MOSAIC framework combines the outputs of multiple inverse models, can improve the overall calibration of the SSL estimate. The approach allows a robot that has learned to calibrate its SSL output in different environments, to select an appropriate set of calibrators as it moves between the different acoustic environments, combining their calibration effort in proportion to how well they are able to calibrate in a particular environment (regardless of which environment those models learned in).

This chapter focuses on whether combining the outputs of multiple models improves the performance of SSL calibration, compared to that using a single model. The performance of the multiple models system was also compared to GCC-PHAT as a popular SSL algorithm, although this latter comparison is of limited value since the system is designed to calibrate the output of an SSL algorithm, rather than perform as an SSL system in its own right. Indeed, the multiple models system proposed in this chapter could have been used to calibrate the output of the GCC-PHAT algorithm itself.

As important as the accuracy of SSL in this thesis (perhaps more so, as a proof of concept) is whether the system can successfully adapt to different environments, having learned in each of those environments, and to what extent the system can adapt to novel environments using existing learned models. The main point is that the system sits alongside an SSL unit, which in principle could use any SSL algorithm. Hence, robustness to background noise, multiple sound sources and so on comes from the underlying SSL mechanism, and the system which is the subject of this chapter calibrates the SSL output for different environments.

Whereas in the previous chapter, the system attempted to identify the context based on the model that gave the lowest error, this is not really the motivation of the system presented in this thesis, although it does point to an interesting potential application in context identification. Identifying the context of course has its uses however, in this chapter, the motivation is how to select a set (because it could be more than one) of models that improves the calibration performance in different acoustic contexts. It should not even matter what SSL algorithm is used- the system should find the best set of models that can improve the performance regardless of the underlying localisation technique used (although changing the SSL algorithm would probably demand re-learning of the models). Also, it should not really matter whether the "correct" model is selected (the function of the system described in Chapter 6 relied on the model that had learned in a particular context being selected); what matters is that the *best set* of models is selected, that is, the combination of models that results in the lowest overall error in azimuth estimation. The input to the system is a SSL estimate, and the purpose of the multiple models calibration system is to "fine-tune" the estimate in the face of errors introduced in a variety of acoustic contexts.

The system that was developed in this chapter is shown in Figure 33. The SSL unit produces an estimate of the sound source azimuth assuming no acoustic distortion or interference. This is analysed into parallel fibre signals (as described in Chapter 4) for each of the adaptive filter models of the cerebellum. Each model produces a calibration signal at its output assuming that it is operating in the context in which it learned to calibrate the SSL output. A copy of the SSL output is summed with each cerebellar output to produce a set of calibrated estimates of the sound source azimuth. Each of these is compared to the ground truth sound source azimuth (when it becomes available, which, in the field, it is envisaged would be via sensory feedback of the sound source position in the environment), and an error generated for each model, which is transformed into the likelihood that each model is the best suited, or is responsible for calibrating the SSL output (the likelihood that each was responsible for the context). Each of these likelihood values is fed into the RE, which uses a softmax function to normalize the likelihoods across models to produce a set of responsibility signals, one for each model. Each responsibility signal is multiplied by a copy of the calibration signal from the corresponding cerebellar model and the set of results is summed to produce an overall calibration signal which is added to a copy of the SSL output to form the system output,

a calibrated azimuth estimate. In this way, all models contribute to the calibration of the SSl estimate, but in proportion to how well each is suited (based on sensory feedback of the ground truth azimuth) to do so. This is the equivalent in the MOSAIC framework of the outputs of the inverse models being modulated by the modules' responsibility and summed to produce an overall motor command.

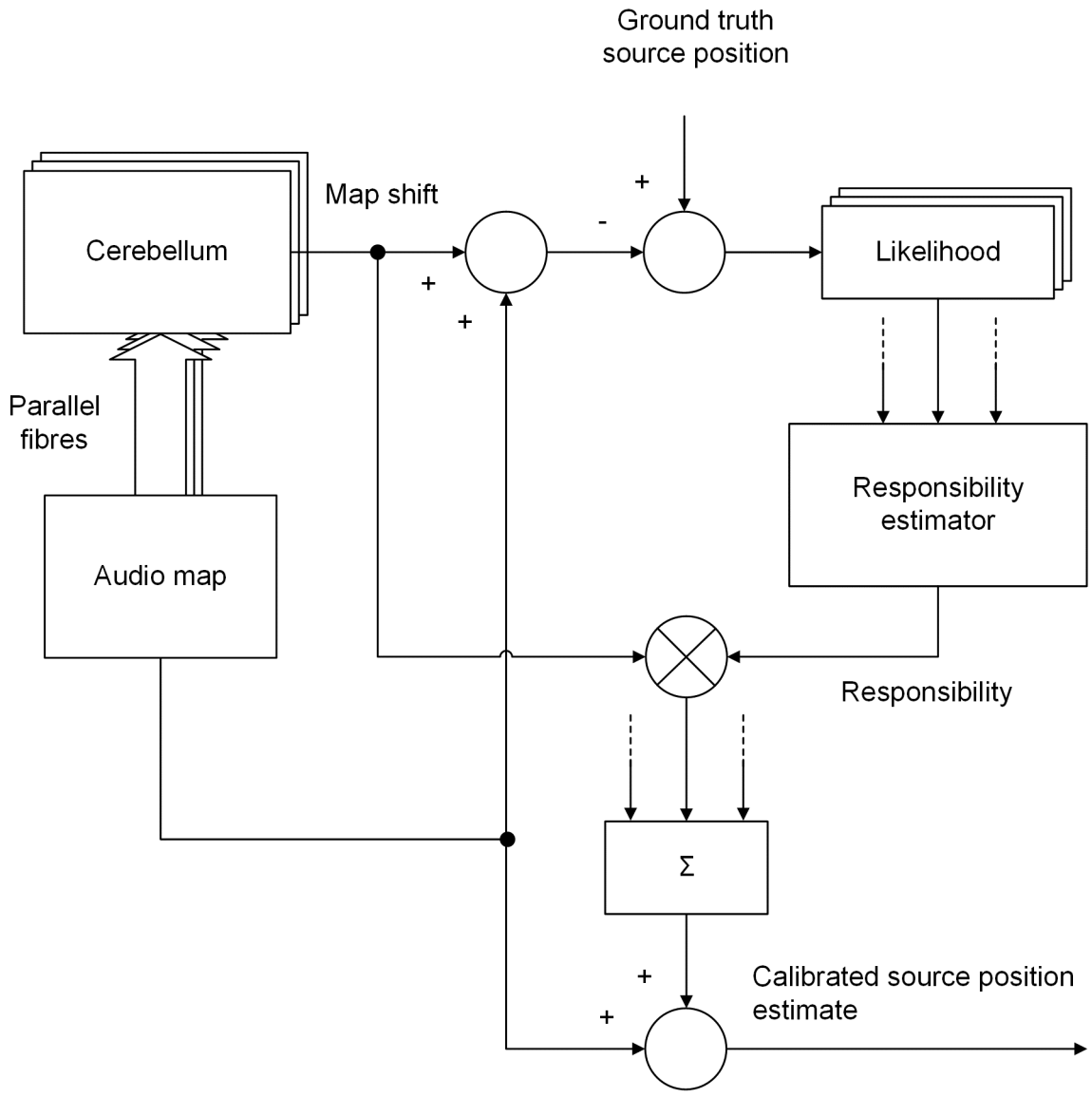


Figure 33. Multiple-models- inspired audio localisation. For a given context, each model provides an estimate of source position. The RE produces a responsibility signal for each model, based on the posterior likelihood calculation. The overall map shift is produced from a summation of model map shifts in proportion to their responsibility. Reprinted from [54]. © 2018 IEEE.

7.1.1 Responsibility estimation

Each model produces a calibration signal based on the underlying SSL estimate which is added individually to that estimate to produce a calibrated azimuth estimate for each model. The estimate derived from each model is compared to the ground truth sound source position. How the ground truth is determined is described in Section 7.2. A robot operating in the field could determine the ground truth through sensory feedback of a different modality, such as vision. This is entirely consistent with the MOSAIC framework in which sensory feedback is used to determine the prediction error of each forward model, after action has taken place or at least commenced. In this case, the robot would orient its vision sensor (if this is the sense being used) toward the estimated sound source (an overall estimate based on the predictions of the models). Although the experimental apparatus was designed to allow such visual input, the ground truth was taken from the odometry of the apparatus used. This was to allow a wide range of experiments to be carried out under similar experimental conditions, using recorded audio. Relying on visual feedback to determine the ground truth would have severely restricted the scope of the investigations. Because the experiments used pre-recorded audio it was possible to immediately know the ground truth without waiting for action or orientation, however this is not how the system would work in the field, where the ground truth would not be available until after orientation, and so this delay of one trial was built into the experiments.

As described in Section 5.2.2.3, in MOSAIC, the prediction error is transformed into a likelihood that the module of which the forward model is a part is *responsible*, that is, the forward model (as it is in the MOSAIC framework) produces the lowest prediction error. Here, again, the system is inspired by MOSAIC rather than being a faithful reproduction. The model itself does not directly estimate the source position (analogous to the next state in MOSAIC) but rather produces a calibration shift to compensate the error introduced into the un-calibrated estimate of source position. This error is normalised, by the RE, across all modules to produce a responsibility signal for each module.

7.1.2 System output

The overall output of the system is a calibrated estimate for the sound source azimuth. Whereas each model's calibration output is individually added to the un-calibrated estimate of azimuth in order to compute the responsibilities, those same calibration shift values are summed, in proportion to each model's responsibility, to produce an overall

calibration shift, which is finally added to the un-calibrated SSL output, to produce a system output which is a calibrated estimate of azimuth.

7.2 Method

7.2.1 Experimental setup

A similar setup as described in 6.2 was used, although the sound source motion control system was redesigned for improved reliability and portability. As explained in section 6.4, the previous setup proved to be unreliable, with frequent adjustment of the sound source locomotion mechanism and its tension being necessary during experiments. The sound source was suspended from a counterweighted tripod-mounted beam (Figure 35). A geared stepper motor (controlled via an auxiliary connection on the PTU) was used to move the beam to place the sound source at various azimuth positions with respect to the robot head. Another advantage of this arrangement is that it is more portable, being more mechanically robust than the arrangement described in Chapter 6 and allowing the possibility of conducting experiments away from the laboratory. Figure 34 shows the robot head.

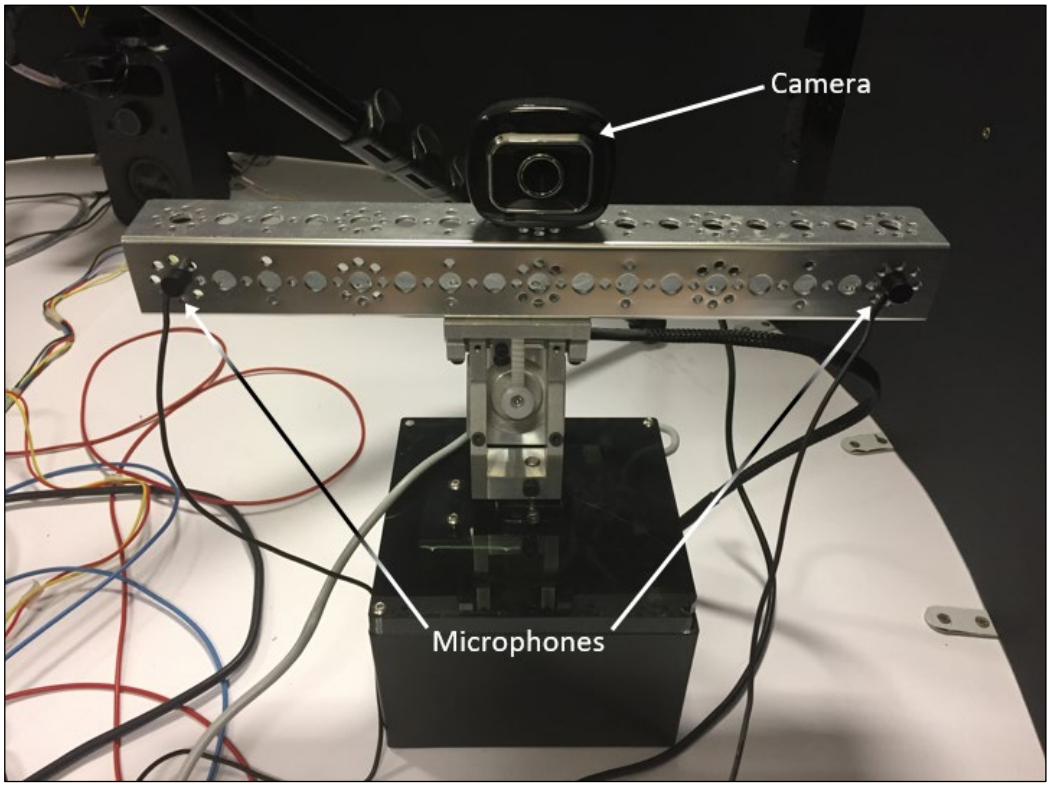


Figure 34. Robot head based on PTU. A horizontal bar with a microphone mounted at each extremity is mounted on a stepper motor to allow the head to orient toward the estimated azimuthal position of a sound source. Reprinted from [54]. © 2018 IEEE.

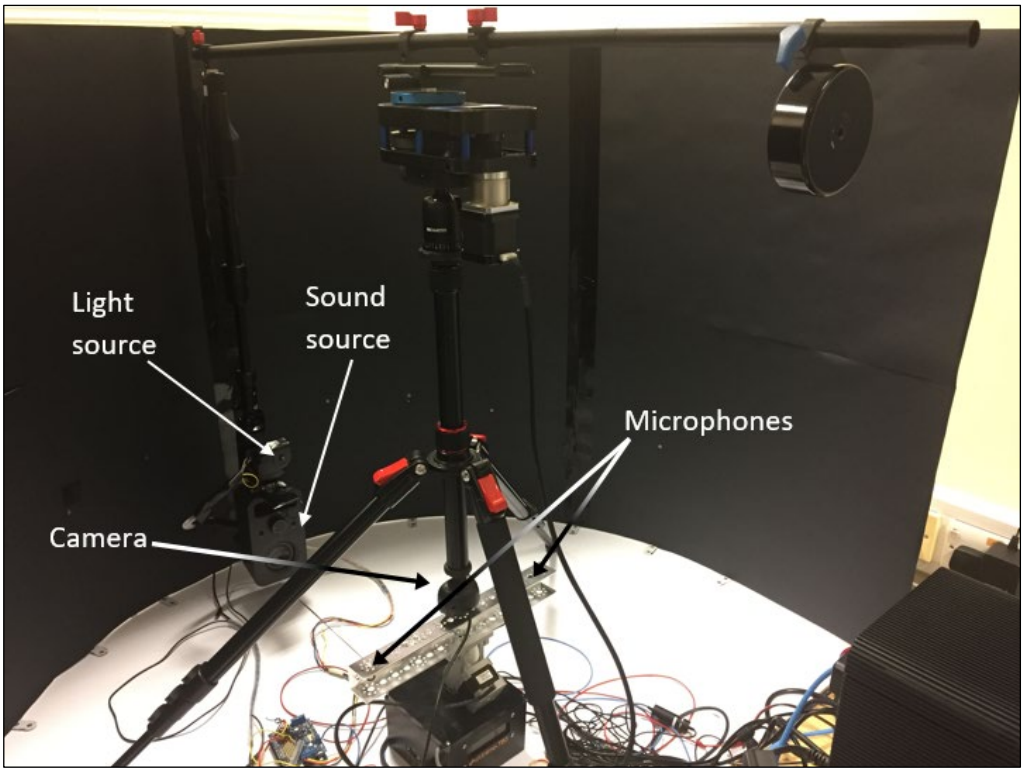


Figure 35. Experimental apparatus using tripod-mounted sound source. The sound source is suspended from a counterweighted beam, which rotates about the axis of the robot head. Reprinted from [54]. © 2018 IEEE.

7.3 Results

Each experiment consisted of a sequence of trials in which he robot head was presented with audio stimulus (a 1 second duration Gaussian noise). The trials were grouped into a sequence of environmental contexts, with 5 trials per context and a randomly selected sound source azimuth in each trial, as though the robot were moving from one environment to another, experiencing 5 audio stimuli in each context, with the azimuth of the sound source in each trial randomly selected (of course, this is a somewhat artificial scenario- the equivalent of the sound source moving about the environment at random; however, this artificial, random selection of sound source azimuth was considered to be more challenging than what would happen in a real scenario and is therefore a valid approach). It is assumed that with a robot operating in the field, a trial would consist of the robot first receiving audio input, then, having made a calibrated SSL estimate, orient its vision (or whatever) sensor toward the sound source, at which point the ground truth could be determined (assuming, of course, that vision is taken to be good enough to provide the ground truth sound source position). Localisation performance was calculated

from 10 runs of each experiment of 15 trials, so that statistics were based on 150 data points.

Due to the posterior nature of the responsibility calculation, in any one trial, the responsibility value used is based on the model's estimation of sound source azimuth during the previous trial. In each trial, an overall calibration signal is generated by summing the individual models' outputs in proportion to their responsibility. In the field, this would be used to orient the robot head toward the sound source in order to determine the ground truth. In this thesis, the ground truth was known from the odometry as explained in Section 7.1.1 and in order to simulate the behaviour of the system in the field was not made available until the next trial. This posterior calculation of the responsibility results in a delay of one trial for the system to respond to a change in context, such as would be caused by the robot moving to a new environment. This is because the ground truth cannot be found until after action has taken place.

7.3.1 Contexts in which the models were trained

7.3.1.1 Multiple models versus best single model

There is a question of whether combining the model outputs in proportion to their responsibilities is better than switching to the single best model in each context (in a similar fashion to Narendra et al. [126]). An experiment was conducted in which the MSE of the combined models was compared to that of the system which switches to the best single model (that showing the lowest error) in each context. Table 2 shows that the performance is better where the model outputs are combined, with a considerably lower MSE than with switched (best single) models.

Table 2. Performance of multiple models versus best single model.

Method	**MSE (degrees^2)**
1. Best single model in each context	12.00
2. Combined models	3.01

7.3.1.2 Multiple models versus a general single model

It is claimed in Section 5.2.1 (following the MOSAIC literature) that a single model would be unable to capture the range of contexts encountered by a robot or organism, and so the performance of the combined models was compared to a single model that had learned in all the contexts encountered. This was seen as a key benchmark against which

to test the performance of the multiple models approach. The single model learned using randomly selected data from across the contexts, rather than being trained in a sequence of contexts in order to avoid the model adapting best to a single context. The model learned over 60 iteration of weight updates in each context. Of course, this is also a somewhat artificial situation as a real robot might visit the three contexts in a more uniform way, spending more time in a context before moving to the next. Also, it is not clear how the sequence of contexts experienced might affect the model's relative performance in those contexts, if the context experience were sequential. In order to compare to an existing benchmark Generalized Cross-Correlation [16] with Phase Transform (GCC-PHAT) was also included as it is a widely used SSL algorithm, although it has to be borne in mind that comparing against an un-calibrated SSL algorithm is not necessarily useful.

Figure 36 shows the responsibility signals of each model. Line charts have been used for clarity although the functions plotted are not continuous, but rather are discrete with integer index values. The plots show that it is the model that has learned in a given context that dominates the responsibility. Where there is sharing of the responsibility, it is mostly with the adjacent model (the third model typically showing zero or near zero responsibility). The result of the ground truth only being available in the next trial (simulating the effect of the system having to wait for the robot head to orient toward the sound source to determine the ground truth) results in a delay of one trial before the system responds to a change in context. This delay in adapting to a new context can result in large errors in SSL calibration until the RE has been able to update the responsibility signal after action, and this was a motivation for going on to investigate the use of the RP in chapter 8. Table 3 shows that the performance of the multiple-models system is better than that of a single model that has learned in all contexts, as well as GCC-PHAT.

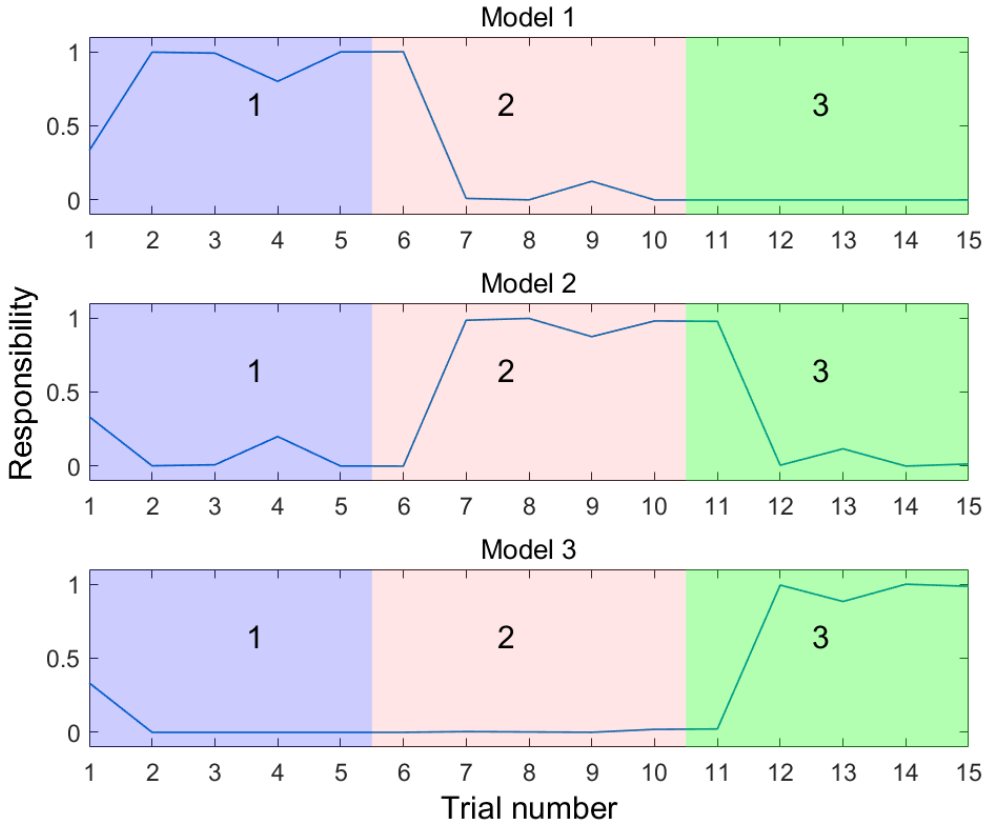


Figure 36. Responsibility signals as the system progresses through the 15 trials. In each trial the system is presented with stimulus of various azimuths in three different contexts, indicated by the coloured regions, labelled with the context number. Context 1 (blue region) is ϕ=90o left; context 2 (red region) is ϕ =0º; context 3 (green region) is ϕ=90º right. Reprinted from [54]. © 2018 IEEE.

Table 3. Localisation performance of multiple models. N=150. Accuracy rate is percent less than 5º absolute error. Adapted from [54]. © 2018 IEEE.

Method	**Accuracy rate**	**MSE (degrees^2)**
1. Single model trained in all contexts	79%	13.5
2. GCC-PHAT	77%	13.6
3. Combined models	92%	5.8
4. Single model trained in novel contexts	88%	11.8
5. GCC-PHAT in novel contexts	86%	10.9
6. Combined models in novel contexts	91%	8.8

As described in section 4.2.3, the experiment was repeated with 81 uniformly selected azimuth values from, and a paired-sample t-test carried out on the calibrated and uncalibrated results, using the Matlab ttest() function. The h value was 1, suggesting that the null hypothesis (that the difference between the means of the calibrated and uncalibrated samples is zero) is rejected, and the improved performance of the calibration model can be treated as significantly better than that of the uncalibrated SSL estimate at the 95% confidence level. The standard deviation of the difference in mean errors was 2.5°. The t-value was -15.7: the negative sign indicates that the calibrated errors were smaller than the uncalibrated errors and the value is not close to zero (the closer to zero, the weaker the case to reject the null hypothesis).

7.3.2 Novel contexts

As discussed in section 5.2.2.3, the MOSAIC literature claims that the framework is able to adapt to novel contexts whose characteristics lie intermediate to two experienced contexts, and this was the motivation for conducting this experiment.

The same 3 models that had learned in contexts associated with values of ϕ of 90° left; 0° and 90° right were tested in 2 new contexts having values of ϕ of 72° left; and 72° right. These fall intermediate, in terms of defining characteristic (assumed to be value of ϕ), to the contexts in which the models have learned.

Figure 37 shows plots of the responsibilities in this experiment. The models that have learned in contexts closest in characteristics to the novel contexts (models 1 and 3) tend to dominate, but less distinctly than in the experiment described in 7.3.1, where models were presented with contexts in which they had learned. There is more sharing between adjacent models, however Table 1 shows that the performance is still an improvement over that of a single model that had learned in the 3 previous contexts, and is similar to that of the GCC-PHAT SSL method. It is slightly worse than that of the models operating in the contexts in which they learned.

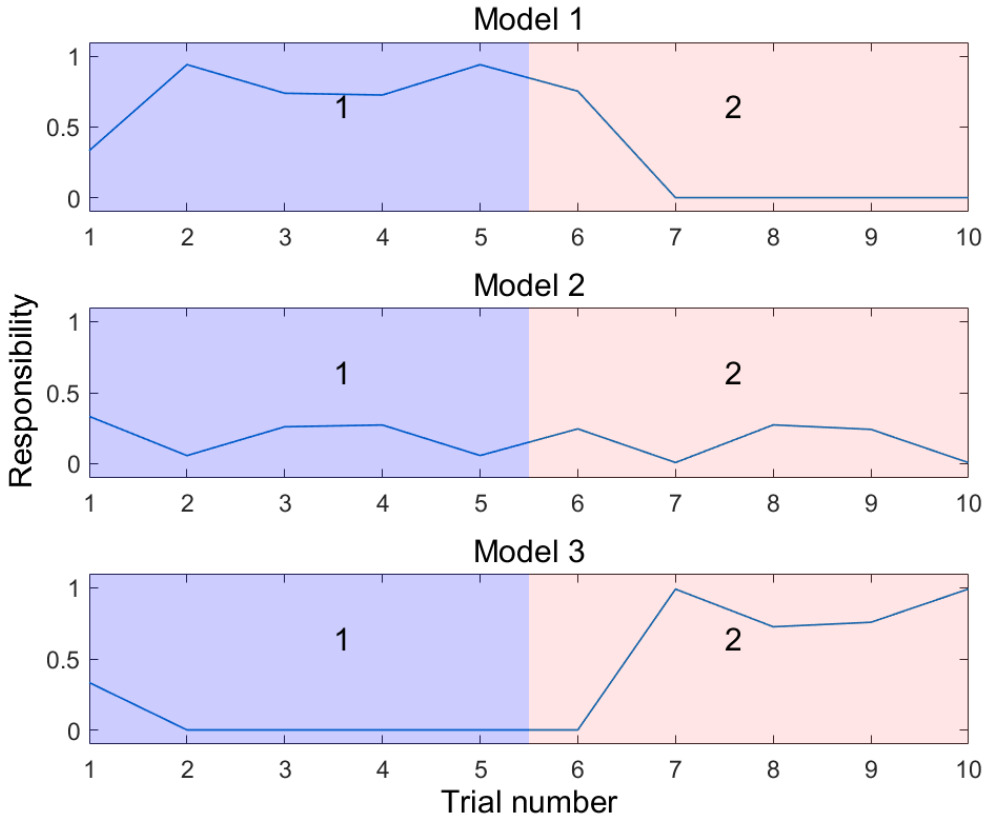


Figure 37. Responsibility signals in novel contexts. In each trial the system is presented with stimulus of various azimuths in two different contexts, indicated by the coloured regions, labelled with the context number. Context 1 (blue region) is ϕ=72^o left; context 2 (red region) is ϕ =72^o. Reprinted from [54]. © 2018 IEEE.

7.4 Chapter summary

This chapter has described a multiple-models-inspired cerebellar calibration system for an SSL algorithm which was able to automatically select an appropriate set of models and combine the outputs of those models, in proportion to their *a-posteriori* determined ability to calibrate the robot's SSL algorithm in different acoustic contexts, to improve the overall SLL estimate in multiple acoustic environments. The performance error of the combined models was better than that of a single model trained in all contexts, in both novel contexts and contexts in which the models had learned. It also outperformed the best performing single model in each context. This represents the most significant result of the thesis, demonstrating the basic idea of multiple models calibration.

Chapter 8 Responsibility prediction

8.1 Introduction

Part of the work in this chapter (specifically Sections 8.2, 8.4, 8.5.5, 8.6.2.2, 8.6.3.2 and 8.6.4, that is, those sections that relate to the *cerebellar* implementation of the RP only) has been published as a full paper to the *Biomimetic and Biohybrid Systems: 8th International Conference, Living Machines 2019* conference. The work in Section V-D was published in [54].

As discussed in Section 6.1 and 7.1.1, the MOSAIC-inspired multiple models system is unable to calculate the responsibilities of the models until after the ground truth sound source location becomes available. For a robot operating in the field, this would be through sensory feedback such as vision, so that the ground truth could only be determined after the robot has oriented its vision sensor(s) toward the estimated sound source position. The MOSAIC framework mitigates this by introducing an RP (see Section 5.2.2.5) which makes a prior prediction of the posterior responsibility (this is the final responsibility output by the RE) of the model to which it is attached, based on contextual input signals. As mentioned in section 7.3.1.2, there is a delay of one trial as the robot moves to a new context before the RE updates responsibility signals which can result in a large error in SSL at the changeover between contexts. This was one motivation for investigating the use of the RP in this chapter. Because the RP is excited by contextual signals, and not feedback in response to action, which has a necessary delay, it can produce a response immediately. This is referred to in MOSAIC as *feedforward* module selection. Also, not covered in the MOSAIC literature, is the possibility in challenging environments that ground truth may not be available at all through sensory feedback, where, for example, vision could become obscured. This is where the RP could play a more prominent role, mitigating performance of the system until the ground truth can be established again.

There are multiple RPs, one for each model, whereas there is one RE, which computes the responsibilities for all models. The RP learns to predict the responsibility of its associated model in a particular context. As with the cerebellar models discussed so far, the RP is pre-trained in this thesis. Ultimately, the RP would learn, or update its learning, alongside the calibration model as the model itself learns. For the purposes of this chapter,

the calibration models are pre-trained, then the RPs trained against the responsibility signals generated as the models act in different environments. The RP in this case needs an already trained model so that it can learn to predict the posterior responsibilities. The actual posterior responsibility outputs of the cerebellar calibrator to which it is attached become the teaching signal for the RP. For convenience, NN implementations of the RP (Section 8.5.3) did not use the posterior responsibilities as teaching signals, which would have required the RP to learn in adaptive mode (rather than the default batch mode in the Matlab NN Toolbox), but instead used RE outputs which had not been produced through a combination of RP output and likelihood values, that is, the normalized likelihood values alone were used, Equation (16). Cerebellar implementations (Section 8.5.5) *did* however use the combined final responsibilities as teaching signals as computed using Equation (17), as was intended in MOSAIC.

8.2 Contextual signals

As mentioned in Section 5.2.2.5, the RP takes contextual signals as input. These are any signals that are derived from the environment and that allow a prior estimate of responsibility, without the need for the ground truth. Hence these signals could be of any form, auditory, visual, tactile, and so on, that allow such a prediction. In this thesis, the audio stream itself is used to derive the contextual input to the responsibility predictor, through the extraction of audio features. There is no reason that vision could not have been used, or, indeed that audio and visual contextual signals could not have been combined, and this may be a fruitful area for future work focusing on audio-visual integration. The rationale for choosing audio, however, was the original motivation to develop a system that could come to rely more on sound where other senses such as vision become unavailable.

In order to train the RP, and for subsequent functioning, a set of features needs to be extracted from the environment, in this case, from the audio stream. There are a number of features that could be used, such as zero crossing rate, energy within auditory filter bank channels and so on. *Acoustic Scene Classification* (ASC) is a recent area of research concerned with the classification of environments based on sounds generated within those environments and is related to CASA (Section 2.3) [136-138]. Features used in ASC have included low-level time- or frequency domain features (e.g. zero crossing rate, spectral roll-off), frequency band energy (including the use of auditory filter banks), cepstral features (see Section 8.4) and spatial features (including ITD and ILD). There is little

literature concerning the application of ASC to robotics. Chu et al. use spectral features of sounds (such as bandwidth and spectral flatness) and zero-crossing rate to classify environments for mobile robot navigation [67].

8.3 RP as part of the overall framework

Figure 38 shows the RP developed in this thesis as part of the overall system. The teaching signal is derived from the RP prediction error which is a comparison between the overall responsibility (for the RP's associated model) and the RP output. As explained in Section 8.1, for convenience, with the NN implementations and for the simulations, the responsibility signal was taken directly as the RE output *without* combination with the RP output. For the cerebellar RP, however, the teaching signal was as shown in Figure 38.

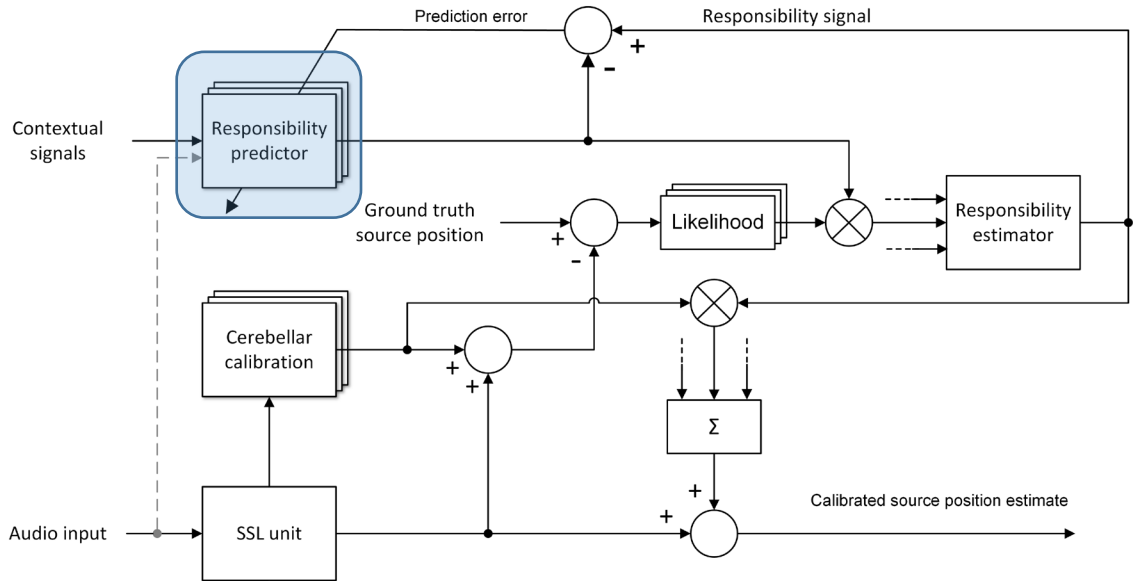


Figure 38. Responsibility Predictor in the context of the overall system. The grey broken line indicates that in this thesis it is the audio stream itself that forms the contextual signal.

8.4 Audio features

Audio features representing the different acoustic contexts were generated using an adaptation of the methods outlined in *Introduction to audio analysis: a MATLAB approach* [139]. This text provides a Matlab library of functions that extract a number of features from an audio signal, the *Audio Analysis Library* (AAL). The library computes 35 features, explained in more detail below, which are zero crossing rate, energy, entropy

of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, 13 Mel-Frequency Cepstrum Coefficient (MFCC) features, harmonic ratio, fundamental frequency and 12 chroma vector features. For each of these features, 6 different statistics are computed (mean, median, standard deviation, standard-by-mean- the ratio of standard deviation to mean, maximum and minimum), resulting in 210 features overall. The audio features are extracted on a frame-by-frame basis (where a frame consists of a number of audio samples, usually 1024; processing individual audio samples would be inefficient and currently impractical, especially for a mobile robot platform, in real-time)- so-called *short-term feature extraction*. Feature statistics are computed over a number of audio frames (so-called *mid-term feature extraction*).

Zero crossing rate and *energy* are time-domain features. Zero crossing rate is the number of times the signal crosses zero (assuming no DC bias of course) divided by the number of samples in the frame, and provides a rough and indirect indication of the frequency content of the signal (so will tend to be higher where higher frequencies dominate, especially "noisy" signals). As will be seen later, this very simple feature on its own turned out to be remarkably successful in signalling the acoustic context. The *energy* is calculated from the sum of the square of the absolute sample amplitudes over the frame (actually, it is the *power* rather than the energy that is computed in the AAL), and tends to be used to differentiate speech from other signals. The *entropy of energy* feature indicates rapid changes in energy level between successive frames. It can be used, for example, to signal the onset of sounds. It is not clear how it would be useful in identifying the environment.

The remaining features listed above are frequency-domain features. The *spectral centroid* feature indicates the central positioning of the frequency spectrum of the signal, while *spectral spread* is a measure of the deviation of frequency components about the centroid. Intuitively, these features ought to convey useful information about the acoustic environment (and indeed turned out to do so as seen later). *Spectral entropy* is the entropy of energy computed in the frequency domain- a measure of the spectral power distribution or "flatness" of a signal, and can be used to differentiate between speech and other sounds. *Spectral flux* is simply the rate of change of (the square of) the amplitude of the spectral components between frames. Again, this feature is useful for discrimination between speech and other sounds. *Spectral roll-off* is an indication of the (near to-) maximum frequency content of the signal (effectively, this is approximately the bandwidth of the

signal). Again, intuitively, this ought to tell us something significant about the acoustic environment. However, for the (somewhat artificial) contexts used in this thesis, this did not emerge as a particularly successful feature. *MFCC* features use a cepstral representation of the signal (where the *Discrete Fourier Transform* (DFT) is taken of the spectrum of a sound) based on a multi-channel filter whose bands are distributed according to the *Mel-scale* (on which frequencies are perceptually equally spaced). MFCCs are heavily used in speech processing, and did not emerge as successful features here. *Chroma vector* is a set of DFT coefficients such that the sound is categorised into classes based on pitch. This feature was developed in the context of music discrimination, and on the face of it will not be particularly relevant here, although, interestingly, one class of the Chroma vector did emerge as a successful feature. *Fundamental frequency* relates to periodic or quasi-periodic signals such as voiced sounds (a normal human voice will possess a fundamental frequency that determines the "pitch" of the voice). In the narrowly defined acoustic contexts here, this is unlikely to be a strong feature, and so it turned out. *Harmonic ratio* is related to fundamental frequency and is computed as the maximum of the auto correlated signal. This is another example of a feature that can be used to discriminate speech and non-speech signals, and interestingly, emerges as a successful feature here.

Just because features did not stand out as particularly successful in this thesis, does not mean that they will not be useful more broadly in more realistic acoustic environments, where a variety of sound types might be present, with a range of characteristics, and the robot might be required to carry out a variety of tasks apart from SSL such as discriminating between human sounds and other types of sound.

8.5 Method

8.5.1 Generation of training data

Training data for the RP was generated using the recorded audio (Section 7.1.1), to generate a set of audio features using the AAL, which were used as training inputs for the RPs developed in this chapter.

For each audio segment in the data set (representing a sound source azimuth and an acoustic context), the trained calibration models together with an RE were used to generate responsibility values, along with likelihood values for the generation of final posterior responsibility targets as described in Section 8.5.5, for the cerebellum based RP.

For the simulations of the RP (Section 8.5.2) and for the NNs (Section 8.5.3), the responsibility signals were used without combination with the RP output. This was done for convenience, since for batch training, the RP output would be required for each training iteration, which was not straightforward using the Matlab NN toolbox. A possible approach here was to batch train the NN repeatedly using a training data set that is increased by one data point after each training iteration so that the RP output could then be used to generate the next training point. This approach was not used although it was attempted in a pilot experiment with some success, which is not reported in the thesis. In the case of the cerebellar RP, it was more straightforward to train the RP with the target signals generated using the RP's own output. This required the generation of a set of target likelihood values using the trained calibration models, as described above, so that these could be combined with the actual RP output during training using a revised form of Equation (16) based on Equation (12):

$$\lambda_i = \frac{\lambda_{pi} e^{-|\theta_t - \theta_i|^2/\sigma^2}}{\sum_{j=1}^{n} \lambda_{pj} e^{-|\theta_t - \theta_j|^2/\sigma^2}} \tag{17}$$

In this way, training data was generated during learning, from the likelihood values, using the partially trained RP.

8.5.2 RP simulation

This work was published as Section V-D in [54]. During the work that was carried out in Chapter 7, the presence of an RP was simulated, to see whether there would be any improvement in performance, i.e. whether it was worth going on to develop an actual RP. This simulation was quite crude, and the actual posterior responsibilities (without combination with the RP itself) were simply used as though an RP were present that could perfectly predict those values. This approach was therefore quite simplistic and not strictly true to MOSAIC, nevertheless it was felt it was worth testing the idea. This was the rational for developing the algorithms and models in this chapter- to test the performance of the system in the presence of an actual implementation of an RP (albeit in software).

8.5.3 Function fitting Neural Network implementation

A function fitting NN was developed using the Matlab *Neural Network Toolbox* `fitnet()` function which creates a network with an input layer of neurons equal in

number to the number of audio features, and a hidden layer with 10 neurons by default (Figure 39). The rationale behind this choice was the ready availability of the NN in Matlab, and that this is the predominant approach used in the MOSAIC literature. The NN was trained to predict the responsibility values of the cerebellar models using offline data as described in Section 8.5.1. To start with, all 210 features generated using the AAL were used as input to the NN.

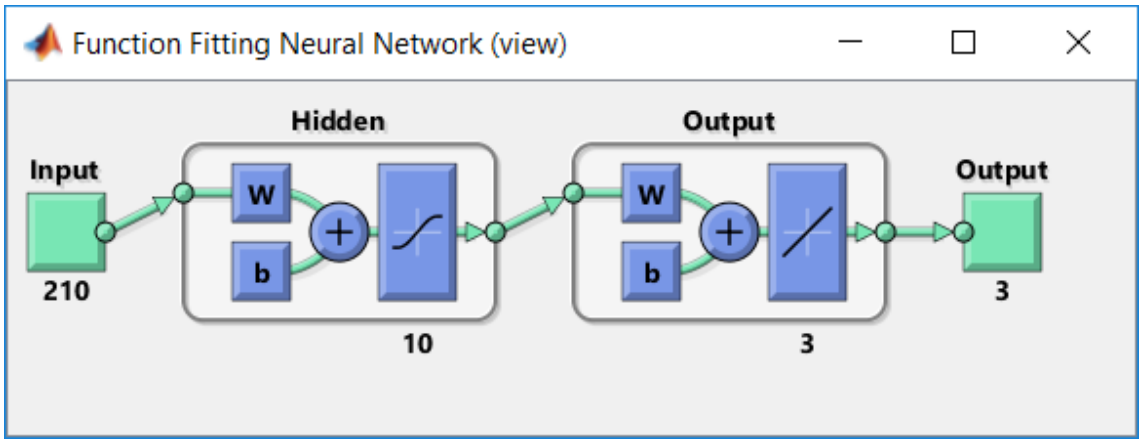


Figure 39. The RP NN structure. As graphically represented by the Matlab NN Toolbox.

A number of training algorithms were tested and *Resilient Backpropagation* was chosen as producing good results with a short training time.

8.5.4 Feature set reduction

Using the 210 different audio features produced using the functions of the AAL is clearly computationally inefficient and indeed may be far from optimal. Too many features can lead to overfitting, reducing the ability to generalise [140]. There are a number of ways to select the most appropriate features to use as input to the Neural Network.

8.5.4.1 Feature Selection

Feature selection is a technique that generates a subset of features. The particular approach used was *sequential feature selection*. There are two types, forward selection in which features are added, and backward in which features are removed. *Sequential forward selection* is a procedure in which an empty feature set is added to, one new feature at a time [141]. As each feature is added to the existing set, a criterion function is evaluated (based on the performance of the network) and if a feature increases performance according to the value returned by the function it is retained in the feature set. The algorithm terminates when the addition of a new feature decreases the

performance of the network. A disadvantage of the approach is that it does not remove features from the set if they subsequently become redundant. It is computationally less demanding than the backward version, *sequential backward selection* (in which the full feature set is the starting point and features are removed one by one), as it operates on a smaller feature set. Preliminary trials showed that the feature set was typically reduced (from the 210 produced by the AAL) to between 3 and 8 features. The Matlab *Statistics and Machine Learning Toolbox* includes a function, `sequentialfs()`. The function takes as arguments predictor and target data (which had been generated from the recorded data set used in Chapter 7) along with a handle to a criterion function that returns a performance measure based on data generated by the algorithm. This criterion function is user generated; it created a function fitting NN (described in Section 8.5) based on the data passed to it by the algorithm and returned the performance of the trained network (MSE).

8.5.4.2 Manual Feature reduction

As part of an investigation into how influential each feature generated by the AAL could be on the performance of the RP, an attempt was made to manually select features. Features were generated from the audio dataset using the AAL and then box plots generated for each feature in each of three contexts (ϕ=-90°, ϕ=0° and ϕ=90°). Each plot was visually inspected and chosen based on how distinct the features were, that is, a feature was rejected if there was considerable overlap of the boxes. As an example Figure 40 shows the box plots for the mean of zero crossing rate. The bottom edge of the boxes indicates the 25th percentile and the top indicates the 75th percentile; the red line indicates the median value; the whiskers indicate the data extremes not considered outliers and the red crosses indicate outliers (the plots were generated using the Matlab *Statistics and Machine Learning Toolbox* `boxplot()` function). Figure 40 shows that there was no overlap between the feature in the three contexts at the 25th and 75th percentiles. This was chosen as the feature used to test the performance of the RP with a single feature input, although there were other candidate features that may have been just as suitable.

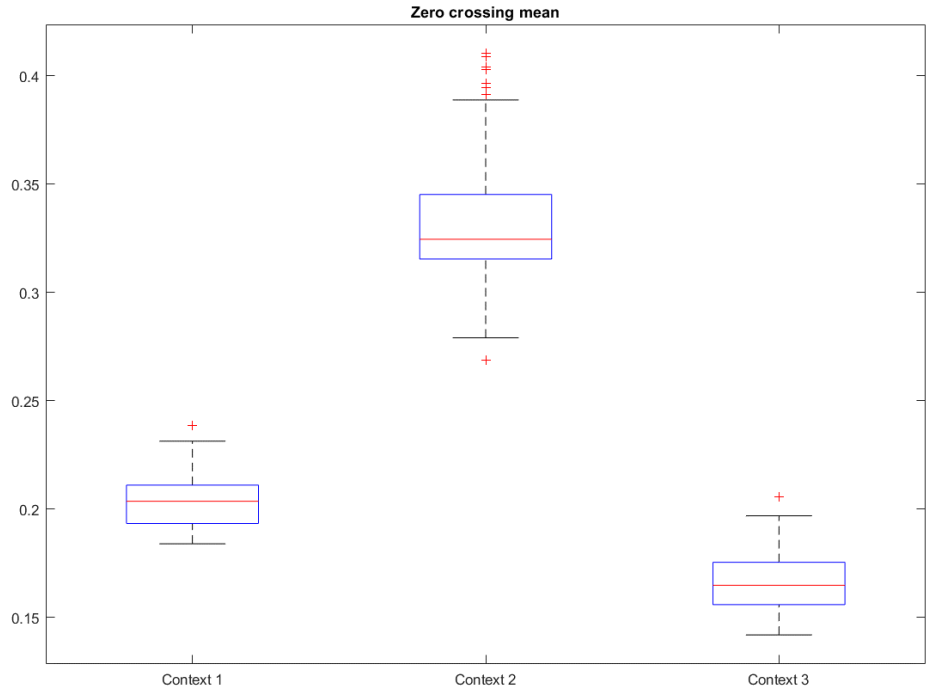


Figure 40. Box plot of the mean of zero crossing rate feature in different contexts.

8.5.5 Cerebellum based RP

A sub- theme of this thesis is to further demonstrate the utility of the "cerebellar chip", where identical cerebellar circuitry can perform multiple functions, which are determined by the context in which it is embedded and its external connectivity.

This connectivity is determined by the way that the cerebellar output (from the Purkinje cell) is used, how the climbing fibre signal (the teaching signal) is derived and how the parallel fibre inputs are derived and connected.

It seems a reasonable approach to investigate a cerebellum-based RP, from two points of view. First, the authors of MOSAIC suggest that the cerebellum is a strong candidate for internal models [120, 132], yet implementations of the RP itself have used more conventional NNs rather than models of the cerebellum [122, 142]. Second, it would seem pragmatic to pursue an approach that makes use of repeated patterns of identical circuitry to perform different functions (in this case, cerebellar calibration and cerebellar RP) in order to facilitate successful migration of the system to a real world implementation on a mobile robot. To this end the same model was used for the RP as for the calibrators. Figure 41 shows the RP based on the adaptive filter model. Features are extracted from

the audio input stream (in this case, just one feature, the mean zero crossing rate was used) and analysed into parallel fibre signals.

The output of the adaptive filter (the Purkinje cell in Figure 41) 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

$$\hat{\lambda}_i = \sum_{i=0}^{n} w_i p_i \tag{18}$$

where $\hat{\lambda}_i$ is the prediction for the *i*th model and *n* is the number of parallel fibres.

The weights are updated in the same way as the calibrators, using Equation (9). As predictions can exceed the valid bounds of 0 and 1 for responsibility, a sigmoid function is included in the output of the RP to limit the output to between 0 and 1.

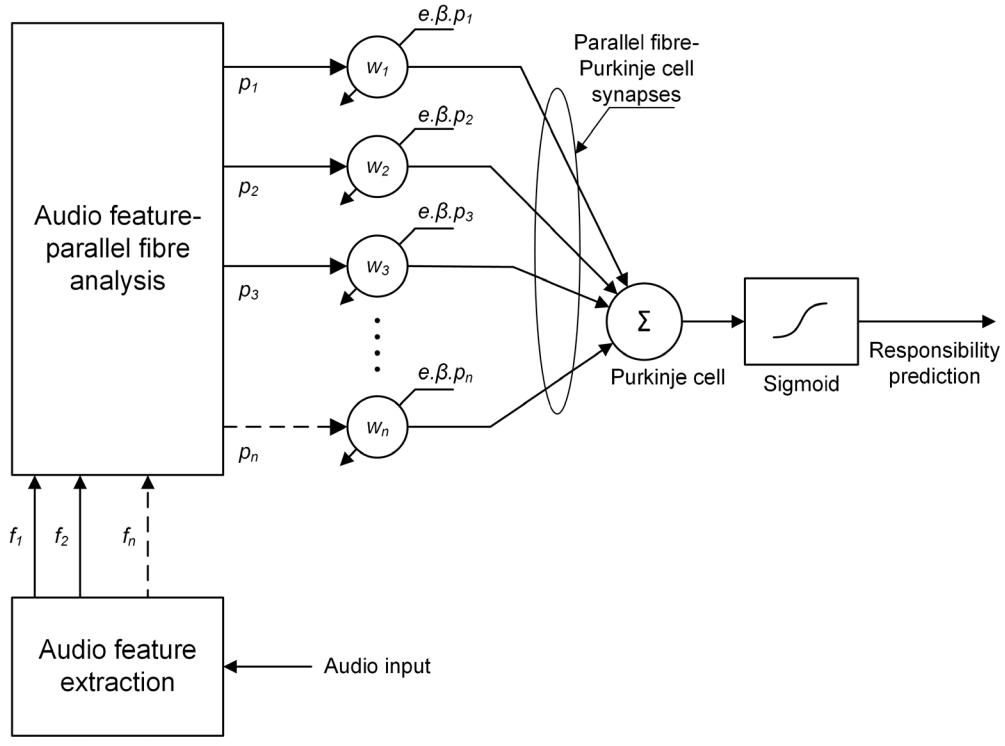


Figure 41. Cerebellar implementation of the responsibility predictor.

Training of the cerebellar RP was somewhat more involved than that of the NN RP. As explained in Section 8.1 and Section 8.5.1, cerebellar calibration models were trained as in previous chapters and used to generate target likelihood values using randomly selected samples of the recorded audio data, with corresponding audio features extracted from the

same audio samples, using the AAL, in each of three different acoustic contexts. Because, in MOSAIC, the RP is trained with the final posterior responsibility value as a teaching signal (that is, a combination of the posterior likelihoods and RP prior), at each training iteration, the partially trained RP was itself used to make a prediction of the responsibility, given the training input, which could then be combined with the target likelihood values that were pre-generated as part of the training data, using Equation (17) to generate a target responsibility. The RP used a learning rate of 16 and was trained over 32000 iterations.

8.5.5.1 Parallel fibres

The basic function of the cerebellum as an adaptive filter relies upon the analysis of the inputs into a number of signal paths and the transmission of those signals via the parallel fibres to the Purkinje cell where they are synthesised into cerebellar output. In the first instance, a parallel fibre configuration close to that used in the calibration model was used, that is, the feature was assigned to a parallel fibre based on its value (much in the same way as that in the calibrating cerebellar model, where it is the azimuthal position of the sound source that determines parallel fibre activity). Each audio feature (in the thesis only one feature was used, chosen through the use of the NN version of the RP as a successful feature). The features were grouped into bins based on value, each bin corresponding to a parallel fibre. With the approach adopted here, 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 extreme of the array. This approach was chosen to be close to the use of the cerebellar model that calibrates the SSL estimate (as well as that used in the precursory study [4]). With that calibration model, the parallel fibre model reflects azimuthal audio activity, so that stimulus toward one end of the limits of the azimuth range would activate parallel fibres toward one extreme while stimulus toward the other extreme of the azimuth range would activate fibre(s) toward the other extreme of the fibre array. Like the calibration model, the RP model therefore has no basis filters. However, investigating basis filter configurations to further process the audio features could a potentially fruitful area of further investigation. In each RP configuration, 50 parallel fibres were used; this number was arrived at through trial and error. A more structural approach to arriving at a number of parallel fibres should be considered in developing the model further.

8.6 Results

The NN implementation of the RP was tested with a variety of feature sets ranging from all 210 features produced by the AAL to just one feature. The cerebellar implementation of the RP used just one feature. With the NN implementation, the neither number of features nor which features were selected appeared to have much impact on the performance of the overall system using the RP.

The cerebellum-based RP typically required larger values of learning rate and learning iterations for satisfactory performance compared to the cerebellar calibration models. Key parameters affecting performance were learning rate, number of iterations, number of parallel fibres and the shape of the sigmoid output function. Typically, a learning rate of 8 to 16 was found necessary for satisfactory performance. Using the standard form of the sigmoid function, several thousand learning iterations were required (typically 16,000 to 32,000), however with careful choice of learning rate and/or sharpness of sigmoid function, this could be reduced to as little as 100 iterations with respectable performance. Compared to the standard NN implementation in Matlab, the cerebellum based RP is somewhat naïve, and it could be that further development of the model would improve performance, with more automated training.

8.6.1 RP simulation

The results of this simulation are shown in Figure 42, where earlier switching of responsibility can be observed between contexts. The localisation performance was improved, with a MSE of 1.5 degrees^2 and 100% accuracy rate (accuracy rate was percentage of estimate errors less than 5^o), compared to 5.8 degrees^2 and 92% respectively for the multiple models without RP (Table 4).

Table 4. Performance of multiple models with simulated RP.

Method	**Accuracy rate**	**MSE (degrees^2)**
1. Combined models *without* RP	92%	5.8
2. Combined models *with* RP	100%	1.5

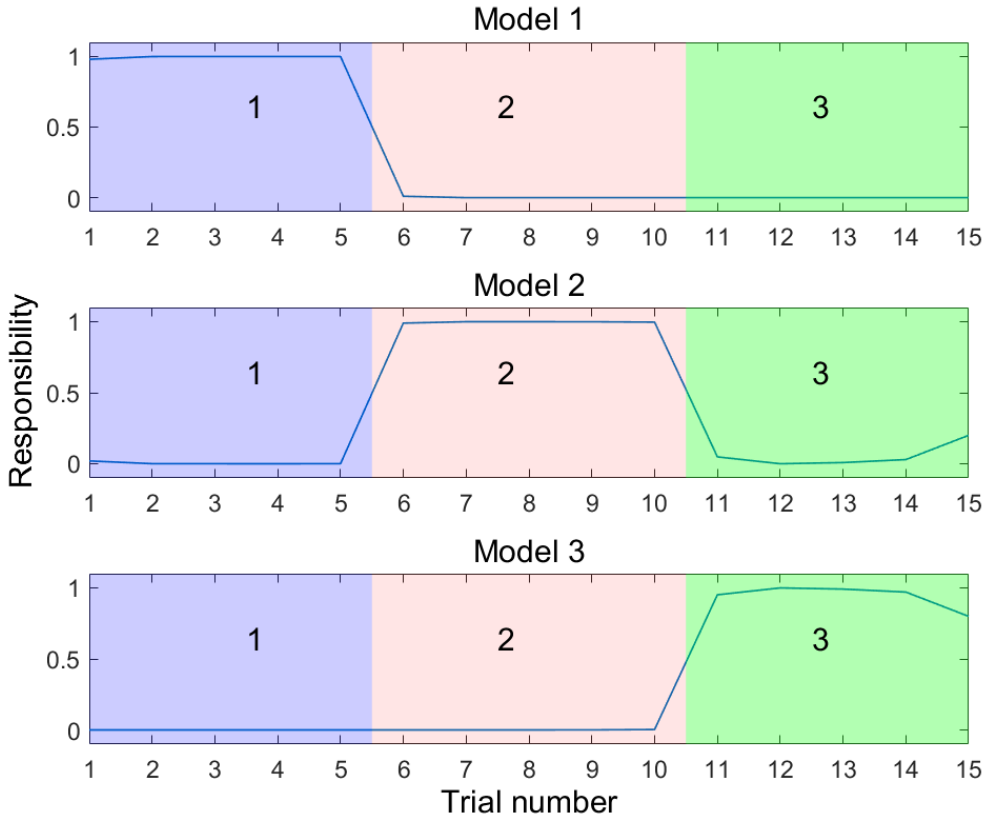


Figure 42. Responsibility signals with simulation of an RP. In each trial the system is presented with stimulus of various azimuths in three different contexts, indicated by the coloured regions, labelled with the context number. Context 1 (blue region) is ϕ=90o left; context 2 (red region) is ϕ =0º; context 3 (green region) is ϕ=90º right. Reprinted from [54]. © 2018 IEEE.

8.6.2 Performance in contexts in which the models had been trained

8.6.2.1 Neural network implementation

The NN worked well for contexts in which the models had been trained. Figure 43 shows the output of the RP with all 210 features from the AAL. Visually, there seems to be a good match between the RP output (orange curve) and the responsibility signals *without* RP (blue curve). The red broken curve is the overall responsibility computed using Equation (17). Because it is a prediction, the RP output is in advance of the RE output. Accuracy rate was 99% (less than 5º error), and MSE 2.1 degrees^2. This compares with results from the simulated RP of 100% accuracy rate and MSE 1.3 degrees^2.

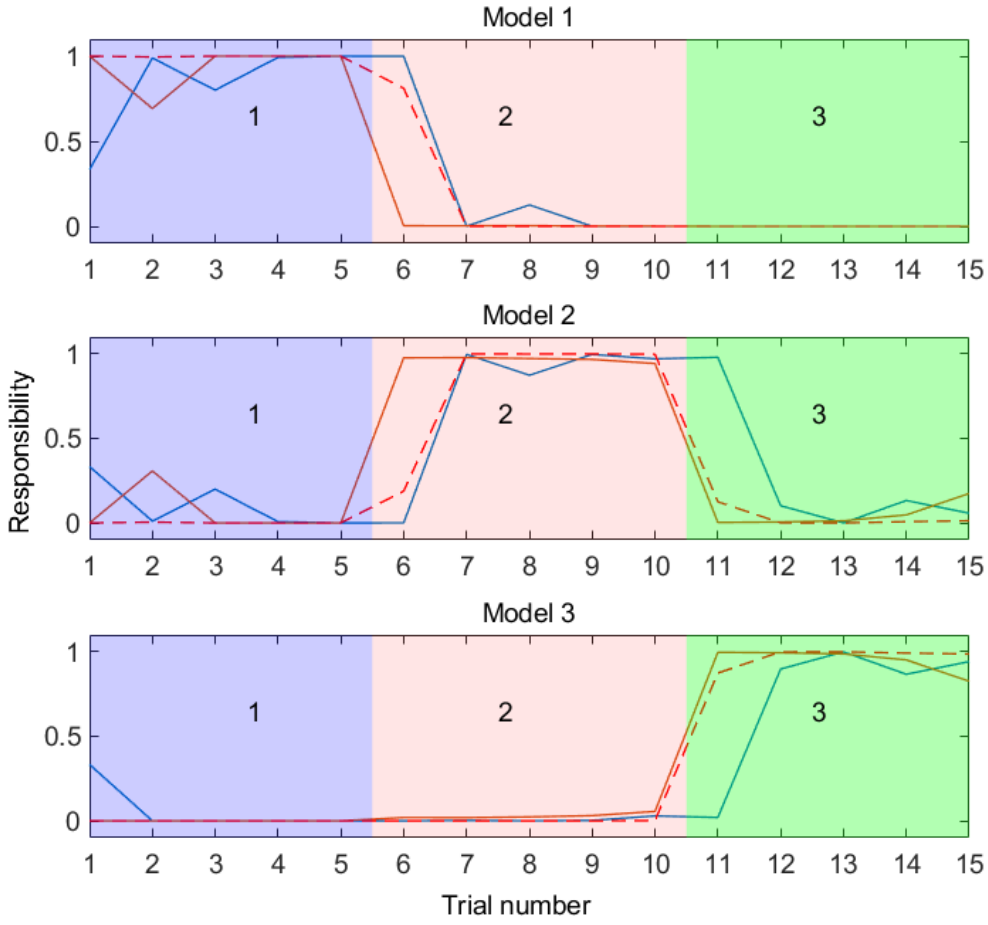


Figure 43. NN RP output using all AAL features. The orange curve is RP output, the blue curve is posterior responsibility, the red broken curve is the combined responsibility

Figure 44 shows the RP performance with just one audio feature (mean zero crossing rate). Even with just this one feature, the prediction visually appears remarkably good. Comparing Figure 43 and Figure 44 it seems that a larger number of features produces a slightly more faithful prediction of the responsibility curve, perhaps capturing more of the variations, although this is unlikely to have a significant impact on performance. This configuration achieved an accuracy rate of 99% and MSE of 2.3 degrees^2, which compares well to the network that used all the AAL features, suggesting that there is little advantage to using such a large number of features (at least, for the contexts used in this thesis- it may be a different matter for more realistic real-world situations).

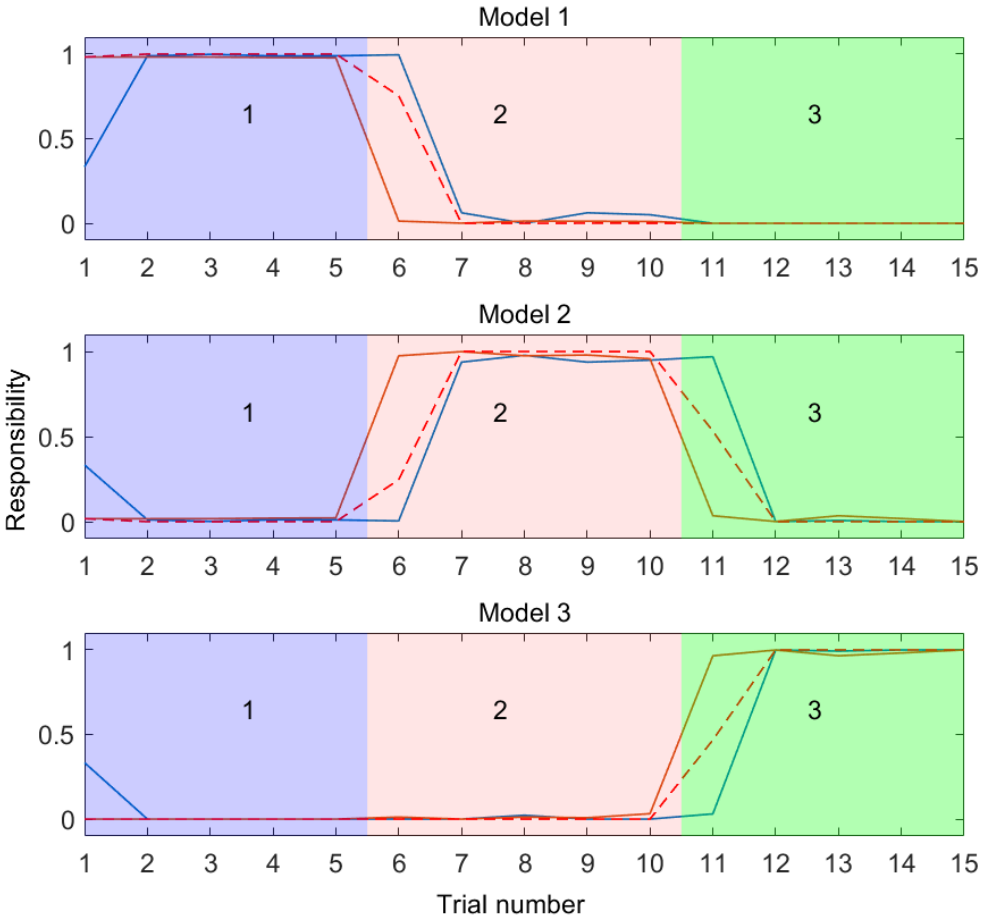


Figure 44. RP output with 1 audio feature (mean zero crossing rate). The orange curve is RP output, the blue curve is posterior responsibility, the red broken curve is the combined responsibility.

Figure 45 shows the RP output using 6 manually selected features as described in Section 8.5.4.2. The features chosen were maximum of energy, standard deviation of energy, standard-by-mean of energy, mean of zero crossing rate, median of zero crossing rate and maximum of zero crossing rate. Accuracy rate was 98% and MSE 2.65 degrees^2. Note that this is a slightly worse performance than that using the single feature alone.

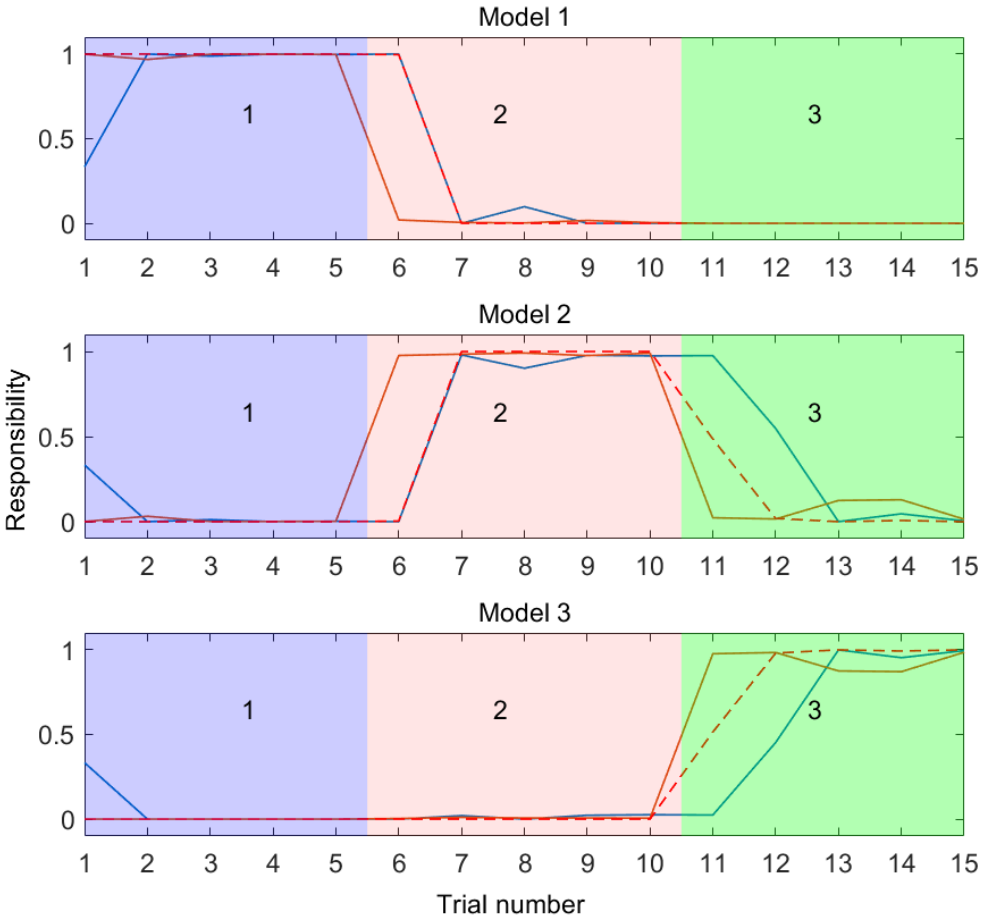


Figure 45. RP output with 6 audio features manually selected. The orange curve is RP output, the blue curve is posterior responsibility, the red broken curve is the combined responsibility.

Figure 46 shows the RP output using 7 features selected using sequential feature selection as described in Section 8.5.4.1. The features selected by the algorithm were mean of spectral centroid, mean of spectral entropy, mean of MFCC feature 11, median of spectral spread, median of harmonic ratio, median of chroma vector feature 10 and median of chroma vector feature 11. Accuracy rate was 99% and MSE 2.27 degrees^2.

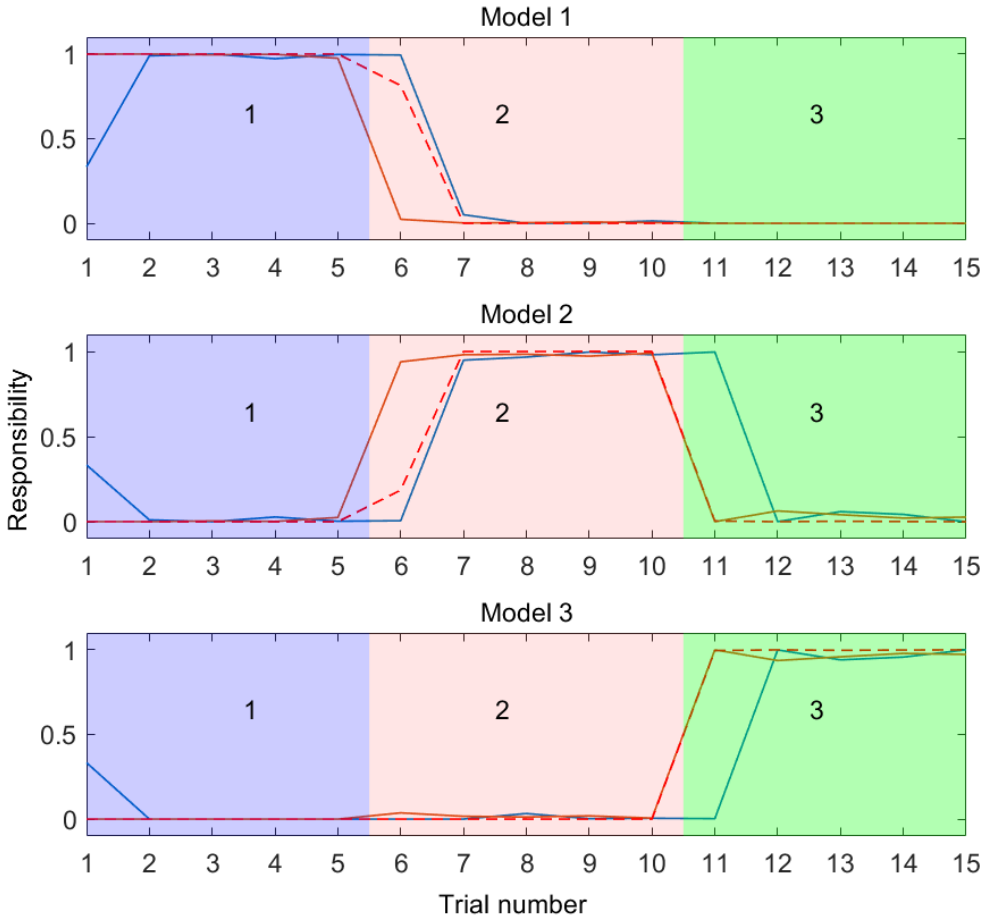


Figure 46. RP output with features selected using sequential feature selection. The orange curve is RP output, the blue curve is posterior responsibility, the red broken curve is the combined responsibility.

8.6.2.2 Cerebellar implementation

The cerebellar implementation of the RP only used one feature, mean zero crossing rate.

Figure 47 shows the responsibility signals of the system as it progressed through trials (each the result of 1 run of the same experiment) in contexts in which the calibration models had been pre-trained. By definition, the RPs had also been trained to predict the responsibilities of the calibration models in the same contexts. The blue curves show the responsibilities *without* RP involvement, that is, the posterior responsibilities alone (generated after the ground truth becomes available), derived directly from the likelihoods using Equation (16). 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 responsibility estimator 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. It can be observed from this figure that the RP output is similar in shape to the RE, 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 (17). The performance of the cerebellar RP was impressive with 100% accuracy rate and MSE 1.1 degrees^2 (over 10 runs).

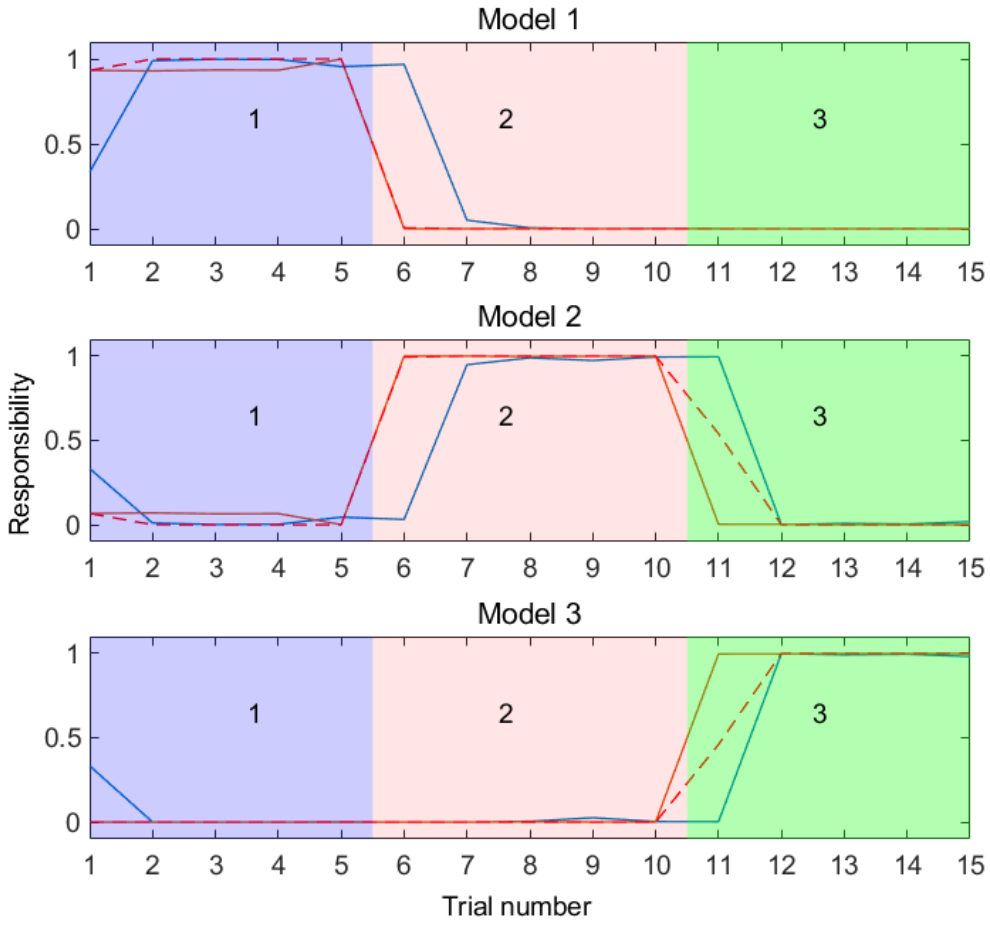


Figure 47. Cerebellar RP output in familiar contexts. The orange curve is RP output, the blue curve is posterior responsibility (without RP), the red broken curve is the overall combined responsibility according to Equation (17).

As described in section 4.2.3, the experiment was repeated with 81 uniformly selected azimuth values from, and a paired-sample t-test carried out on the calibrated results with and without cerebellar RP, using the Matlab ttest() function. The h value was 1, suggesting that the null hypothesis (that the difference between the means of the samples

is zero) is rejected, and the improved performance of the multiple models with RP can be treated as significantly better than that of the multiple models without RP at the 95% confidence level. However, the t-value was smaller than that elsewhere in the thesis at -2.7, suggesting that although the improvement is significant, the case for being confident of this is somewhat weaker than that of the calibration method versus other techniques. This is perhaps to be expected given that the thesis claims a significant improvement using the multiple models alone.

8.6.3 Performance in novel contexts

8.6.3.1 Neural network implementation

Using all 210 features produced by the AAL produces fairly poor results, in terms of distinctiveness of dividing up the experience (Figure 48).

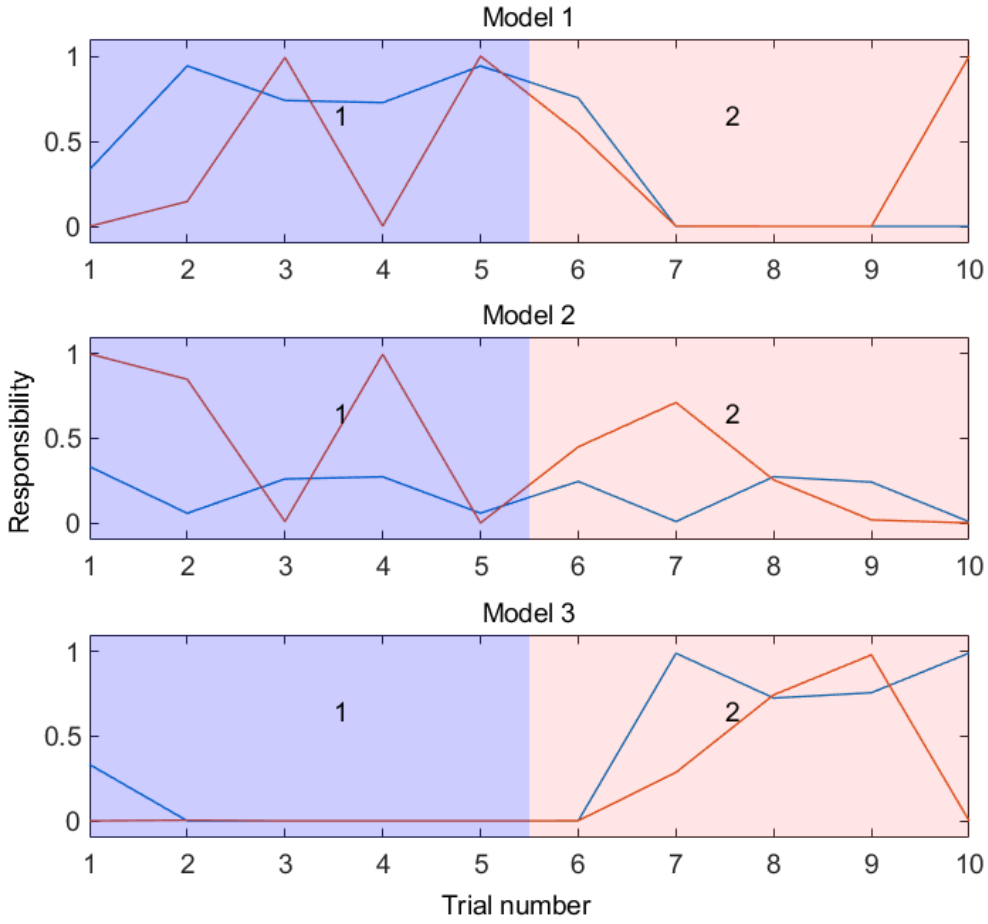


Figure 48. NN RP output in novel contexts using all AAL features.

Figure 49 shows the RP output in novel contexts with a single feature (mean zero crossing rate) as input to the RP and shows a reasonable performance visually. This might be expected, perhaps due to overfitting with a large number of features, as explained in Section 8.5.4.

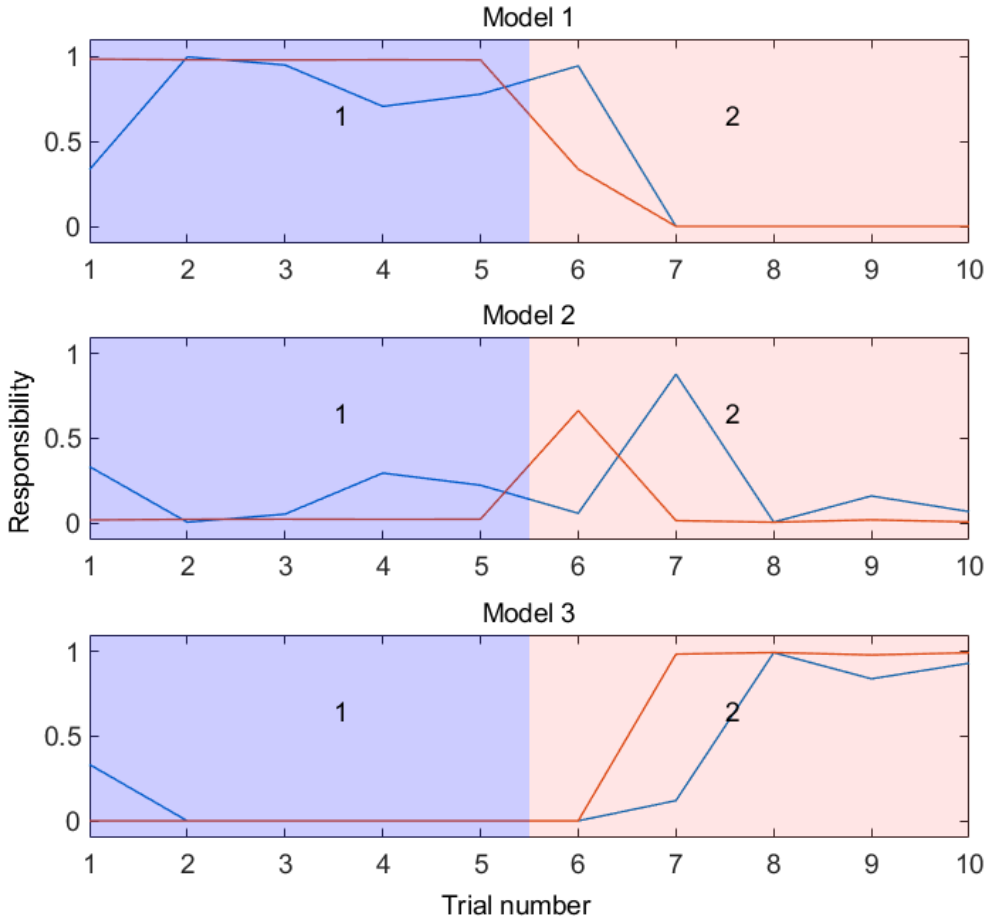


Figure 49. NN RP output in novel contexts using one feature.

8.6.3.2 Cerebellar implementation

The performance of the cerebellar implementation was comparable with that of the NN implementation in novel contexts. One has to be careful in making a comparison, however, as the cerebellar implementation uses the final responsibility signal, which is a combination of transformed posterior likelihoods and RP prior, as a teaching signal, whereas the NN implementation used a simplified version without RP involvement. Figure 50 shows the responsibility signals of the responsibility estimator and the RPs. As in Chapter 7, the system without the RP is able to generalize to the novel contexts, and

the RPs also appear to be able to generalize (or, rather, predict the generalization of the calibration models) quite well. The RP output shows earlier switching of responsibility. However, in this case, the overall responsibility shown by the red broken curve more closely follows the responsibility without RP involvement rather than the RP output itself. The accuracy rate was 95% and the MSE was 5.7 degrees^2, which is an improvement on the system without RP (90% and 9.3 degrees^2 respectively, using the same data set).

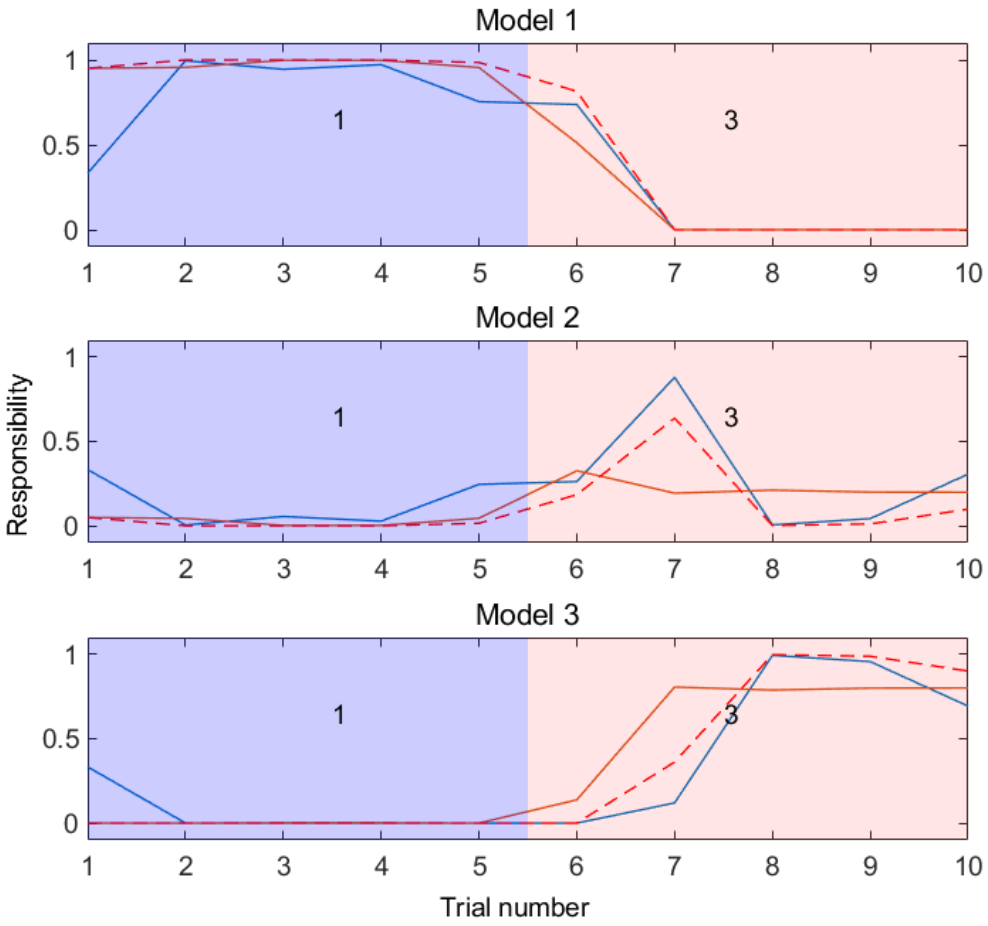


Figure 50. Cerebellar RP output in novel contexts.

8.6.4 Misclassification of the context by the RP

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 [122]. In that work, RPs were trained by presenting a visual pattern corresponding to a context. After training, an incorrect pattern/pairing was presented, and a performance error observed as an inappropriate module was selected a-priori, but then corrected a-posteriori as the RE output dominated.

In this section of the thesis, RPs (based on the cerebellar implementation as described in Section 8.5.5) were trained against their models in each context, however, in the post-training trials, in context 2, the RPs alone were presented with audio stimulus for context 3 instead of context 2. The calibration models however, were presented with the correct audio from context 2. Figure 51 shows that the RPs do indeed mis-classify the context, as would be expected. Because they are presented with the audio from context 3 during context 2, during that context, the RP for model 3 predicts dominance in context 2 as well as context 3 as shown by the orange curve. The output of the RP for model 2 remains low throughout the trials. The RE (blue curve) does correct the overall responsibility when the ground truth becomes available, shown by the red broken curve, which is the overall responsibility, closely following the blue curve during context 2 from trial 7. Similarly, the RE dominates the overall responsibility for model 3, overriding the erroneous high value of the RP in this context. Of course, this relies on the ground truth always becoming available through sensory feedback, and, as mentioned elsewhere in this thesis, this may not always be the case.

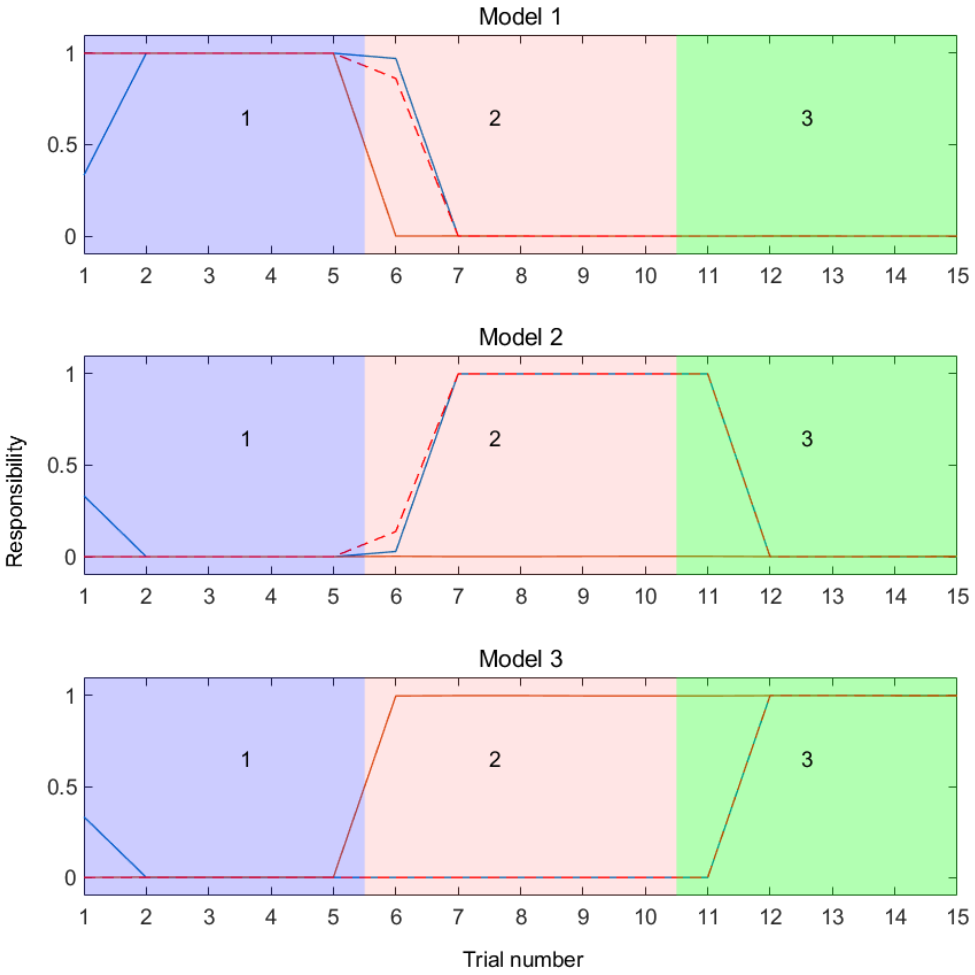


Figure 51. RE posterior correction of RP error.

8.7 Chapter summary

The RP is able to successfully predict the responsibility values of the cerebellar SSL calibration models in the acoustic contexts presented, using features extracted from the audio stream, especially in contexts in which the cerebellar calibrators had been trained. Performance of the NN RP was poor in novel contexts, but was much improved with a reduced feature set. The RP based on the adaptive filter model of the cerebellum performed quite well, better than the NN version with many features, but perhaps not quite as well as the NN version with just one feature. A direct comparison between the NN and cerebellum implementations may not be useful- they were trained in quite different ways and the NN implementations in Matlab are somewhat more sophisticated than the cerebellar RPs developed in this thesis, which might benefit from further development.

An RP that can successfully predict the posterior responsibilities of the proposed system will allow the system to update the responsibilities of the models before the ground truth becomes available, reducing the performance error where the context changes but the RE has not yet updated the responsibilities to reflect this. The number of features chosen seems to have little effect on the overall performance of the RP (this was only tested with the NN implementation), except in novel contexts, and indeed an RP using a single feature, for example the mean zero crossing rate of the audio stream, performs well. This may not be the case, however, in different and more challenging real-world environments, where perhaps more, and perhaps different, features will be required.

Last, it was shown that where the RP itself makes an error, misclassifying the context and pre-selecting an inappropriate set of models, the RE corrects for this *a-posteriori*, when the ground truth becomes available, as is claimed in MOSAIC. However, this relies on the ground truth becoming available through sensory feedback, which may not always be the case.

The main significance of this chapter is that it demonstrates the mitigation of missing ground truth, on which MOSAIC relies. Although MOSAIC includes an RP to mitigate errors due to a change in context before the ground truth becomes available through sensory feedback, the literature does not deal with the possibility of the ground truth availability being disrupted, a real possibility in challenging environments, and this is addressed in this chapter. Relying on the RP alone in these circumstances allows the system to perform in a comparable to fashion to when the ground truth is available.

Chapter 9 De-novo learning of multiple models

9.1 Introduction

The ultimate goal is for a robot operating in the field to be able to move from environment to environment, adjusting its SSL calibration in environments it has previously encountered, but adapting to new environments as it comes across them. As discussed in Section 5.2.2.3, a certain amount of adaptation to novel contexts is inherent in the MOSAIC system, where the characteristics of those environments fall intermediate to those of environments the system has experienced, and this has been displayed in this MOSAIC-inspired system (Sections 7.3.2 and 8.6.3), using existing trained models. There must come a point, however, where this form of adaptation is insufficient to deal with completely new contexts. As discussed in Section 5.2.2.3, a weakness of the MOSAIC system is that generalisation to novel contexts only works with interpolation, and the system cannot extrapolate to novel contexts whose characteristics lie outside the bounds of those of the contexts in which the models learned. In these situations, it seems reasonable to assume that the system would need to learn a new model when confronted with such a situation. Alternatively, if the characteristics of the novel environment are not too different from those of a known environment, perhaps the corresponding models would adapt through learning. This seems to be the approach suggested in the MOSAIC literature, in which the learning of modules is modulated by the responsibility signals so that it is the modules best suited to the new environment that can re-learn in the new environment. The MOSAIC literature does not appear to address how *new* models would be generated, and there is a question as to how we know when this would happen, or alternatively, when existing models would re-learn to adapt to the novel environment, and how this would be implemented. The thesis side-steps this question by concentrating on the de-novo case, and this was the approach used by the authors of MOSAIC [122], where the number of models was pre-ordained to match the number of contexts to be experienced. As such, this approach is not actually self-organisation, as the models pre-exist, one for each context to be experienced. Models started from randomly selected initial conditions (model parameters, i.e. neural network weights), and the models then learned competitively, with the teaching signal of each model modulated its responsibility. Of course, this still leaves the question of how we would determine the required number of models beforehand, in a real situation, we would want to “release” a

naïve robot, with no models at all into a variety of environments and have generate models as required and have its multiple models calibration system divide up the experience between the different models as they learn. This question of how the required number of models would be automatically determined from a tabula rasa state, is left to future work (Section 11.3.8), but seems a very interesting problem. Escobar-Juárez et al. [143] used a *Self-Organising Map* (SOM) based architecture to generate models, and this may be a potential are to investigate.

The motivation in the de-novo case though is to be able to present a set of untrained models with audio stimuli from the same number of contexts as there are models and have each model emerge as the dominant model in a particular environment (insofar as it displays the greatest responsibility of all the models in that context). In this sense, the although system is not learning from a truly tabula rasa state, as the models already exist, they *are* untrained, whereas previously in the thesis, they have been trained each in their own context.

The cerebellar models used in previous chapters were initialised with zero weights, an approach inherited from the software adapted from the precursory project (Bella), and one which reflects the "silent synapses" of the cerebellum. However, this would be unsuitable for de-novo training of the models, as all models would learn in an identical fashion (and indeed, pilot trials, not reported in the thesis, showed this to be the case), so that all models would behave in an identical way and there would be no advantage over using a single model. Another issue is that it is anticipated that the models may not equally divide up the experience and that one model may dominate in each new context learned, so that a means to bias the learning may need to be used. This could easily be done by manually adjusting the responsibility signals during training (in fact, it could be that the system, when coming across a new context, which could be signalled by a large overall error, would generate a new model and enable the teaching signal for this model only). In the event, this domination by one model in all contexts did not happen, as shown in Section 9.3.

The approach taken was to randomly initialise the weights of a pre-existing set of calibrator adaptive filter models of the cerebellum. The range of the weight values was chosen somewhat arbitrarily and seemed to have little effect on the outcome of the learning. The exception to this appeared to be where contexts were similar in

characteristics and one model dominated in each context, but that this was mitigated by varying the range of initial weights (Section 9.3.1, page 123, and in particular Figure 56 and Figure 57).

By modulating the teaching signal of each model with its responsibility signal as described in Section 5.2.2.3, it is envisaged that cerebellar models will competitively learn in each context. Pilot experiments showed that the models were still able to learn to calibrate the audio map with weights initialised in this way, rather than with zero weights as in previous chapters.

A model that happens to display a smaller error during training will receive a larger share of the teaching signal, so that Equation (9) can be re-written as

$$\Delta w_i = -\lambda_i \beta e p_i \tag{19}$$

where λ_i is the responsibility as calculated using Equation (16).

Of critical importance to the success of this approach is the value of σ in Equation (16) used to calculate the responsibility. Too large a value and there would be little discrimination between the models and all models would learn equally; too small and the value of responsibility calculation easily exceeds the precision of the machine running the algorithm as the exponential term in Equation (16) becomes very small. It was found that a considerably smaller value of σ was required for successful dividing up of experience (a value of 0.003 produced good results compared to a value of 2 when the system was in operation). It should be noted that a clear distinction is made here between learning and operation of the models. During operation, the teaching signal is removed, although the MOSAIC system does allow for adaptation during operation as described in Section 5.2.2.3. The approach to de-novo learning taken in MOSAIC [122] is to initialise the models to particular starting points.

9.2 Method

The initialised models were presented with a sequence of contexts and allowed to train concurrently in each context. On each iteration, the error of each model was computed and multiplied by the models' current responsibility value before the weights of each

model were updated. The same number of models were trained as the number of contexts presented which was varied from 2 contexts (and 2 models) to 4 contexts (and 4 models).

9.3 Results

9.3.1 Two contexts with two models

Figure 52 shows plots of the parallel fibre-Purkinje cell weights after training during context 1, corresponding to a value of ϕ of -90^o (this was the first context presented). In this context, the sound source is facing to the left from the point of view of the robot head. This will tend to cause an error in SSL estimation to the left, which by convention is a negative shift in the estimated azimuth, which in turn would require a positive compensatory shift to be produced by the models. The weight values are distance around the azimuthal arc, from centre (straight ahead of the robot) in metres. Very little learning has taken place in model 1 while model 2 appears to have dominated the learning.

Figure 53 shows the weights after learning in context 2, corresponding to a value of ϕ of +90^o (which was presented after context 1). The weights of model 2 appear mostly unchanged while model 1 has now dominated the learning. Here, the sign of the weights is opposite to that of model 2 which has learned in context 1. This is to be expected as the error introduced by the sound source facing in the opposite direction to that in context 1 will tend be of the opposite sign (that is, having the sound source facing right with respect to the robot head will tend to introduce an azimuth error towards the right). Also, the weight index represents a region on the azimuthal arc: index numbers of 5 or greater represent parallel fibre inputs for sounds coming from the right while those of value 4 or lower represent sounds coming from the left. For this reason, model 2's peak in weights occurs at around index number 5 while model 1's is around index number 4 (we would not expect a large difference between the two as the parallel fibres are set up to accept sounds from a 360^o azimuth range, so that an index of 1 corresponds to a sound at azimuth of -180^o while an index of 8 represents a sound at azimuth +180^o).

Figure 54 shows the responsibility signals of the two models, post learning. The experience has been distinctively divided up between the two models, with model 2 dominating in context 1 and model 1 dominating in context 2.

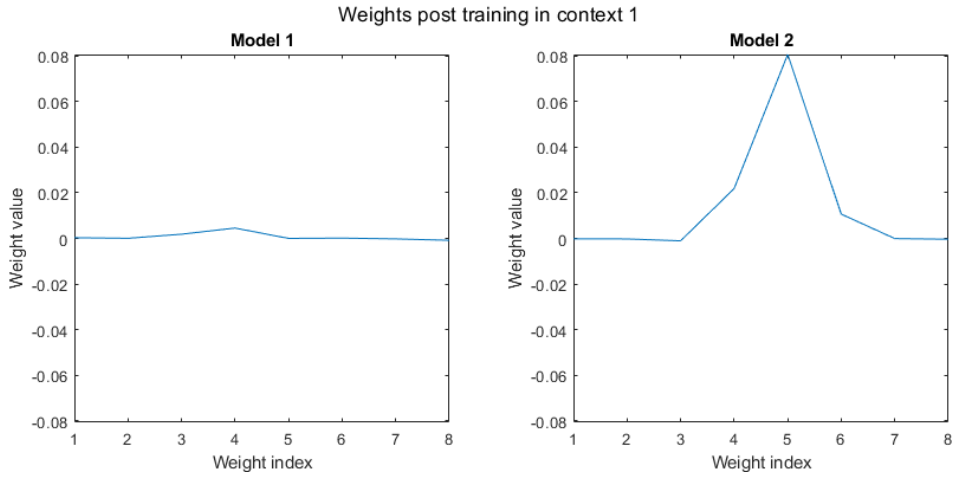


Figure 52. Cerebellar weights after training in context 1 of two contexts. ϕ_1=-90º.

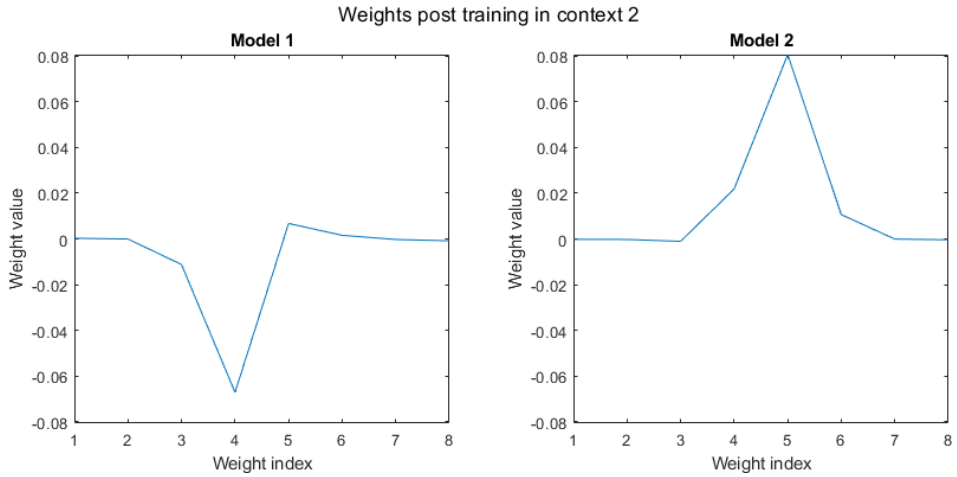


Figure 53. Cerebellar weights after training in context 2 of two contexts. ϕ_2=90º

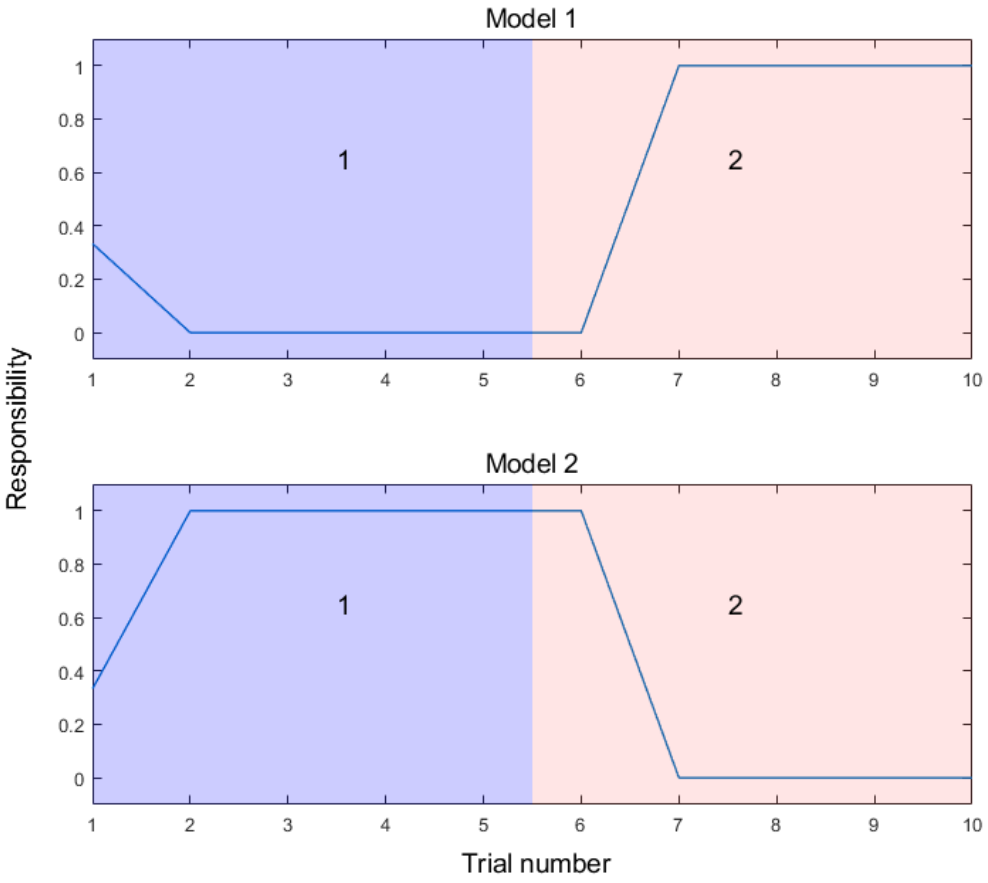


Figure 54. Responsibility signals post- de-novo training. ϕ_1=-90º, ϕ_2=90º.

It should be noted however that the two contexts chosen are quite distinct in characteristics (value of ϕ). Choosing contexts that are closer in characteristics leads to less distinct dividing up of the experience. For example, choosing a value of ϕ for context 2 of 0º instead of +90º leads to a very slightly less distinct sharing of responsibilities (Figure 55), although the difference is hardly noticeable, but with closer characteristics (ϕ=-45º in context 2), one model appears to dominate in both contexts (Figure 56). In this case model 1 has learned in context1, and then appears to have adapted in context 2. However, it seems that judicious selection of parameters (in this case, choosing a much smaller range for the initialised weights, on the order of 10^{-7} rather than 10^{-3} used in all other experiments where weights were randomly initialised) results in some distinctive dividing up of experience (Figure 57). This points to successful de-novo learning being sensitive to a number of factors: the distinctiveness of the contexts experienced, the choice of the value of σ, and the range of the initialised weights.

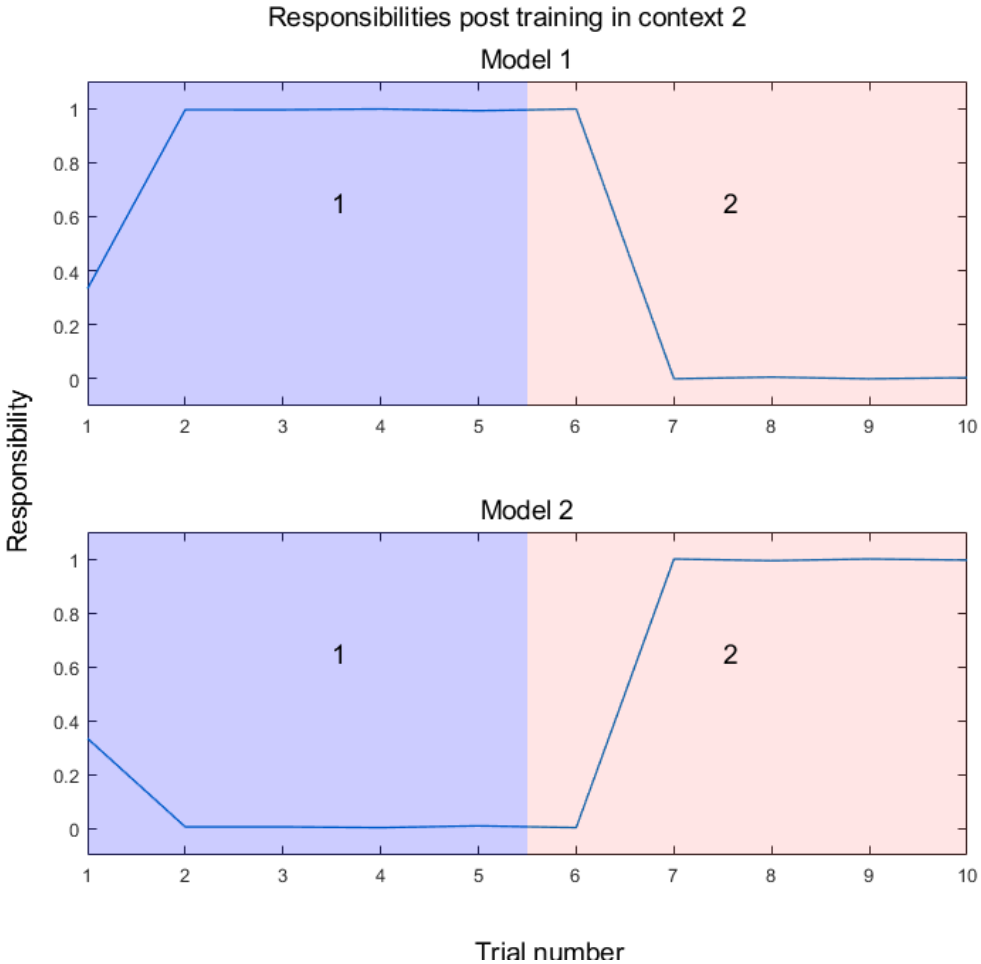


Figure 55. Responsibility signals after training in slightly less distinct contexts. ϕ_1=-90º, ϕ2=0º.

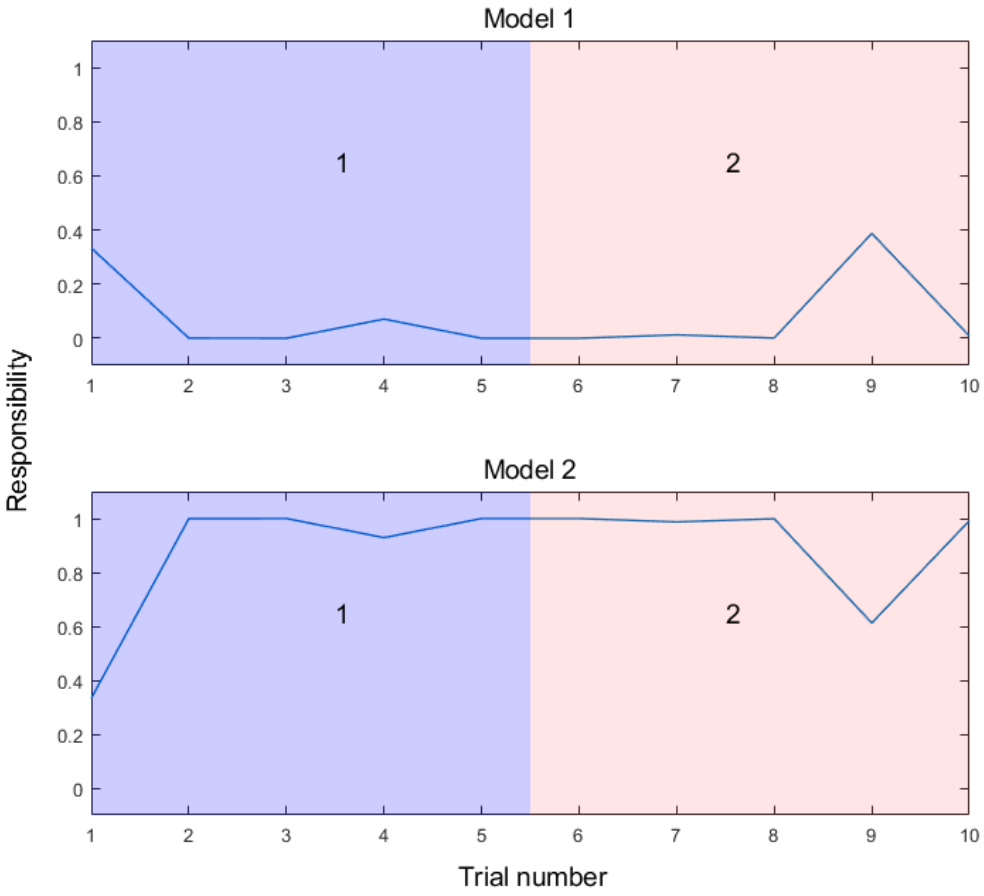


Figure 56. Responsibility signals after training in less distinct contexts. ϕ_1*=-90º,* ϕ_2*=-45º.*

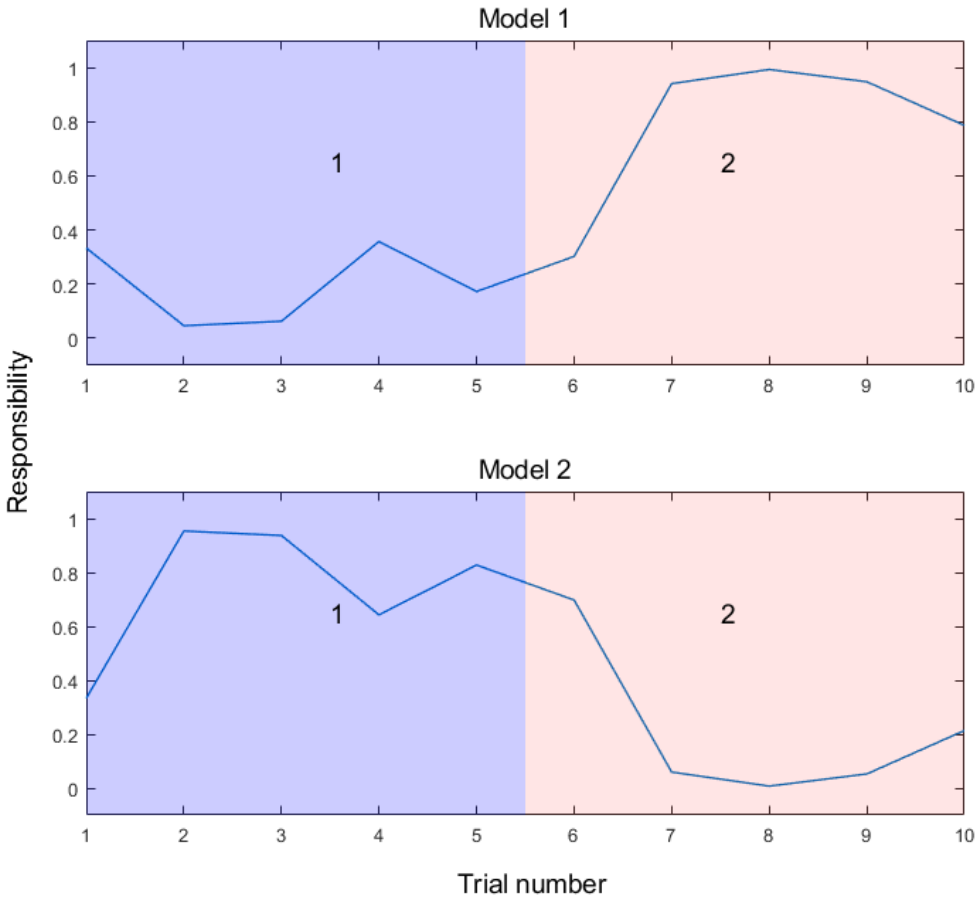


Figure 57. Responsibility signals after training in less distinct contexts with lower range of initial weights. ϕ_1=-90º, ϕ_2=-45º.

9.3.2 Three contexts with three models

Figure 58 shows the responsibility signals of three models that have learned from an initialised state. The experience has been quite distinctively divided up between the three models, but less so than in the case of two models and two contexts. In this case model 2 dominates in context 1, model 3 in context 2 and model 1 in context 3. Contexts 1 and 3 are the same as contexts 1 and 2 respectively in Section 9.3.1, and context 2 is zero azimuth, that is, sound source directly facing the robot head. So, in this experiment, the previous two contexts were retained and a third, intermediate context added. Interestingly, in this new context, in which the sound source directly faces the robot head, model 1 appears to dominate by default rather than by learning, with very little learning having taken place compared to the other two models as shown in Figure 59.

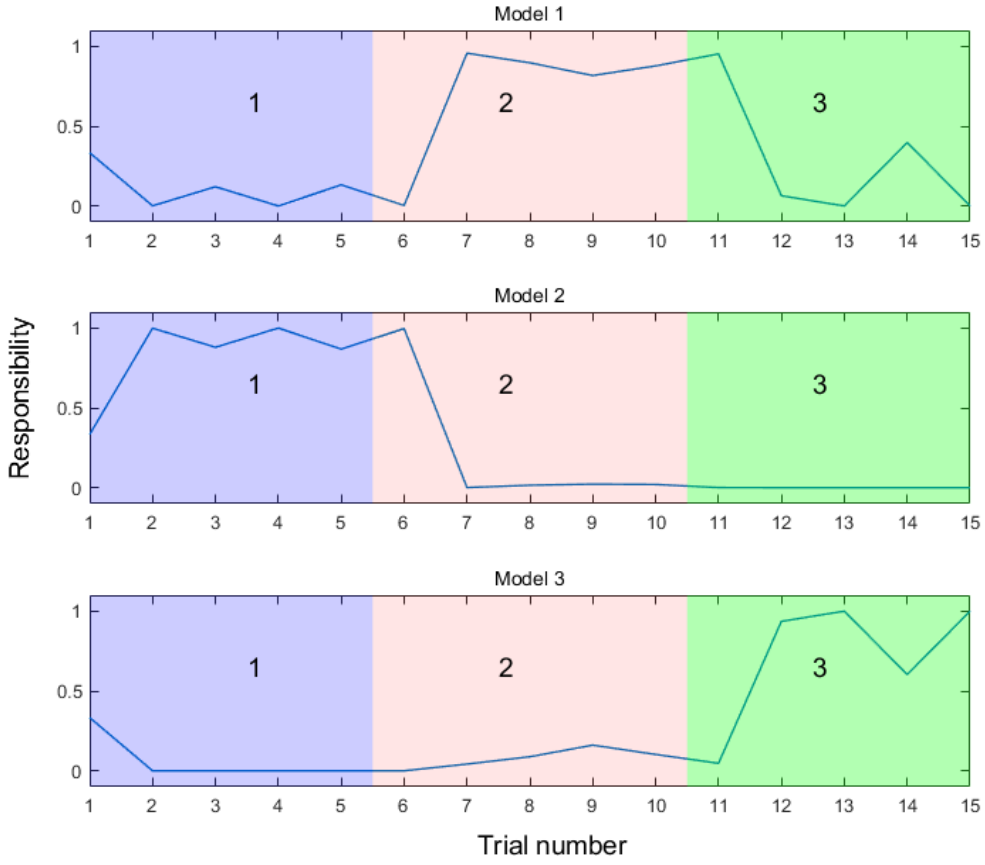


Figure 58. Responsibility signals post de-novo learning: 3 models. ϕ_1=-90º, ϕ_2=0º ϕ_3=90º.

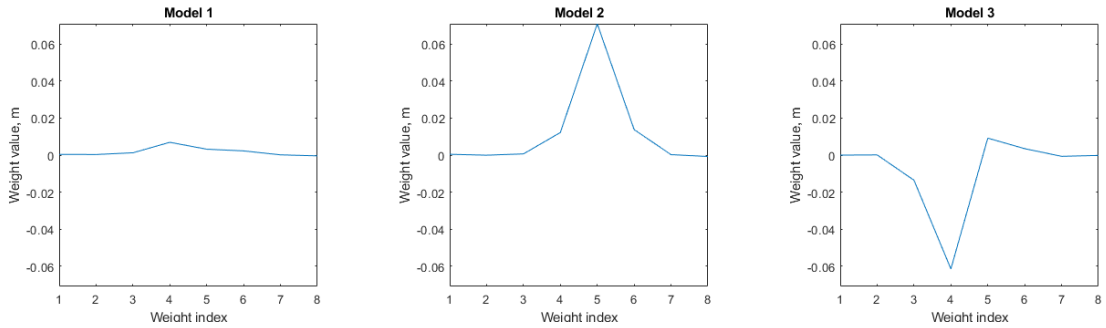


Figure 59. Cerebellar weights post de-novo learning: 3 models.

9.3.3 RP learning post- de-novo learning

This was the most straightforward manifestation of the "complete" system, that is, a self-organising system that includes RPs. In simulations of MOSAIC [122] the RP does not appear to have learned alongside the models but learned against pre-trained models. For one experiment, this approach was adopted, such that the RPs learned against models after the latter had self-organised, using the following algorithm:

```
for each context
begin
     train all models
end

for each context
begin
     train all RPs
end
```

In this situation, the models learned in each of contexts 1, 2 and 3 in turn, and then the RPs learned against their models in the same sequence of contexts. Figure 60 shows that the RPs were able to learn the responsibilities of the self-organised models. Accuracy rate was 98% (MSE 4.0 dregrees^2) compared to 92% (MSE 6.3 degrees^2) for the models alone. RPs were trained with 100 iterations. The value of this result seems a little limited however, as one might expect from the results of Chapter 8, that the RP would successfully learn against its model, regardless of whether that model was manually trained or self-organised.

In another experiment, the RPs were trained against models after the latter had trained in each context. So, for example, models would learn in context 1, then the RPs would train against their models, before the models moved on to learning in context 2, using the following algorithm:

```
for each context
begin
     train all models
     train all RPs
end
```

It was felt that this was somewhat closer to a situation where the RP learns immediately alongside (or rather, immediately following, as model learning must precede RP learning) its model. However, it is not clear from the MOSAIC literature that RPs would learn in this way. In [122] RPs were trained on pre-trained models. Another question that is not

clear from the MOSAIC literature is whether the learning of models is modulated by the responsibility with or without the influence of the RP. In Tidemann et al. it does appear that learning of the models includes the RP output [144]. In this thesis models learned using the responsibility signal without RP influence. Again, the value of this results seems rather limited and it seems to point to the need for further work in this area.

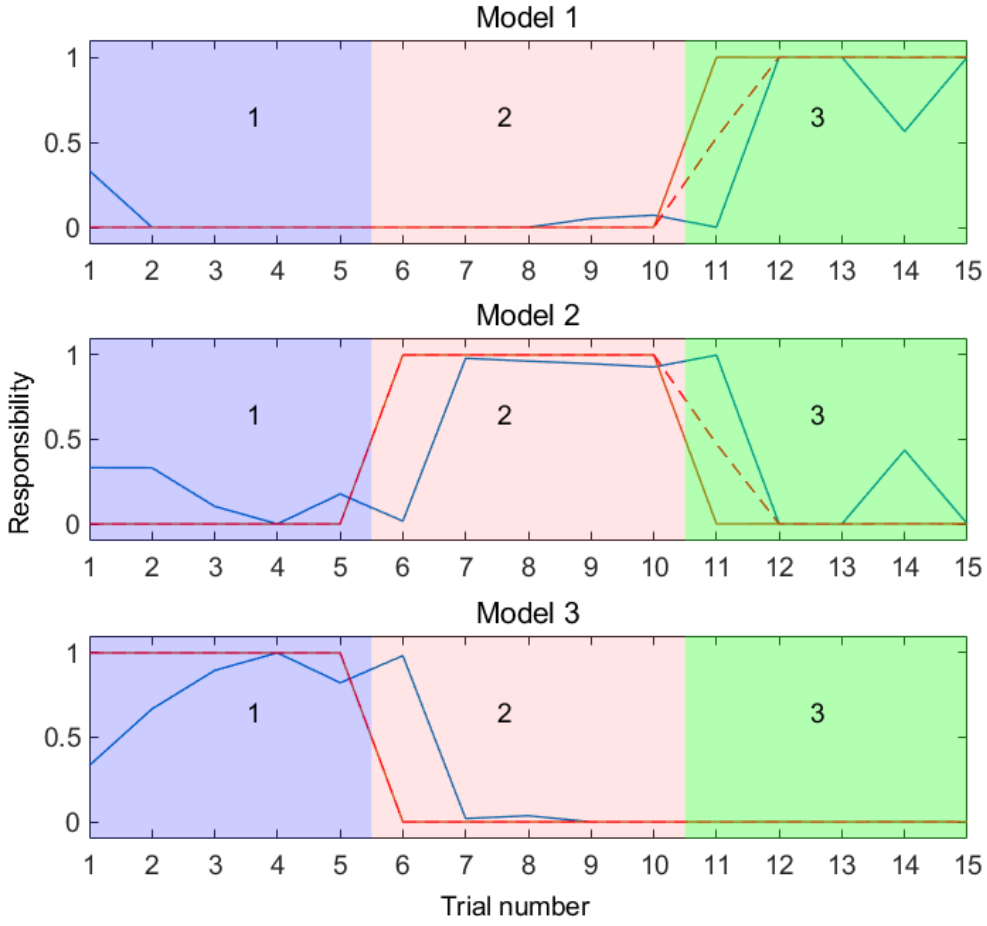


Figure 60. Responsibility signals including RP, post de-novo learning.

9.4 Chapter summary

This chapter has shown that the proposed system has the potential to self-organise, with some limitations, in particular that this is from a de-novo state, with a pre-ordained number of models, so that it is not true self organisation. Although this addresses the third research question, *Can a multiple-models inspired audio calibration system self-organise?* this was included almost as an afterthought to satisfy the author's curiosity about this aspect of the MOSAIC framework, and yet, this has possibly highlighted the

most fruitful and important area of future work, especially in terms of how the RP would learn *alongside* its calibrator model, rather than post learning of that model, and this could be significant in other areas that draw on a multiple models approach, not just robot audition. In the literature, de-novo learning in MOSAIC based systems has been carried out under quite constrained conditions, and often with difficulty, and compared to this benchmark, the system developed here has performed remarkably well. De-novo learning worked very well (in terms of the dividing-up of experience) with two models (provided that the presented contexts were not too close in characteristics), however, the distinctiveness with which the experience is divided up appears to decrease with the number of models/contexts. However, the SSL calibration performance does not appear to deteriorate with the number of models (this is dealt with in Section 10.3.2).

The chapter has also shown that the RP can learn, to a limited extent, alongside the calibration models from a de-novo state. This has not been addressed at all in the MOSAIC literature, where RPs were trained against their models post-de-novo learning of those models. Initially, this was the approach taken in the thesis, so that the calibrator models were first allowed to train competitively (in *all* acoustic contexts), and then the RPs allowed to train against their calibrator model. The system performance was somewhat improved in the presence of the RPs. A (somewhat limited) move towards having the RPs learn alongside the models was attempted, in that the RPs were trained against their models after those models had trained in a particular context. That is, in each context experienced, the calibration models were trained and then the RPs trained against them, before moving on to the next context and training of the calibration models competitively in that context. In general, this led to an improvement in performance in the presence of the trained RP, although this was variable. It was particularly difficult to achieve distinctive training of the RPs in this situation; in most cases the RPs generated roughly equal responsibilities to each other.

This chapter set out to show that the proposed system had the potential to learn from a de-novo state. Although this potential has been demonstrated, there are a number of questions raised that remain unanswered. The thesis uses the same number of models as expected contexts (as is done in the MOSAIC literature), however this is a somewhat artificial scenario- a robot operating in the field with little human intervention ought to be able initialise itself in any set of environments without *a-priori* knowledge of what the number of environments will be, and hence, how many models will be required. There is

still the question of how we know when the existing models should adapt, that is, re-learn, and when a new model should be instantiated, although the investigation of SOMs may prove fruitful here. A particular issue that is not well answered here, and also is not addressed at all in the MOSAIC literature, is how RPs would learn *alongside* their models either from a de-novo or from a tabula rasa state.

Chapter 10 Towards real world environments: bringing it together

10.1 Introduction

Parts of the work in this chapter have been published in *IEEE Robotics and Automation Letters* (specifically Section 10.2, Section 10.3.1 and Section 10.4.1) and other parts have been submitted to *Biomimetic and Biohybrid Systems: 8th International Conference, Living Machines 2019* (specifically Section 10.4.2).

The system proposed in this thesis has been tested under quite constrained conditions, and there is a question of how well the system would perform in real-world, unstructured conditions. The problem with real world environments is the presence of background noise, other sound sources and reverberation, and the changing nature of acoustic environments.

10.2 Performance in domestic environments

This work was published as Section V-G in [54]. The experiments conducted so far in the thesis have been carried out under quite constrained, controlled conditions, in an office/laboratory environment (background sound pressure level of around 30dBA), but with the difference between contexts produced by varying the angle ϕ of the sound source on its vertical axis. However, ultimately this system needs to be of practical use in a robot operating in the field, which means coping with real world, unstructured environments. To this end, some limited experiments were carried out in a domestic household situation. The key factors here as far as SSL is concerned are reverberation and interference from sound sources in the environment other than the one of interest. As discussed in Chapter 2, such environments pose a challenge to SSL schemes, especially the basic ITD scheme used in this thesis.

The same apparatus as used for the work in Chapter 7 was used, and indeed, the tripod mounted motion control system was developed so that experiments could be conducted in real situations away from the office or laboratory.

The experiments were conducted in a domestic kitchen/diner with a hard floor and little furniture, providing a challenging acoustic environment with significant levels of reverberation. The experiment was conducted in different locations in the room, and at different distances between sound source and microphones. The maximum distance was

constrained due to limitations of the apparatus, so that a 1m distance to source was the maximum practical distance that could be used (as the tripod mounted motion control system needed to be counterbalanced, the experiment needed at least a 2.5m diameter clear area which was a challenge in a domestic situation). The dimensions of the room were 3.9m x 3.1m.

Trials were conducted in two contexts, each in the same room: in the middle of the room, with a distance to source of 1m and sound source angle ϕ set to 90º right, and in the corner of the room, with a distance to source of 0.5m and ϕ set to 135º left. Cerebellar models were trained with 100 iterations due to the more challenging conditions. Although performance was poorer than in more constrained conditions, the system still outperformed the single model trained in all contexts as well as the GCC-PHAT algorithm (Table 5).

Table 5. Localisation performance in domestic contexts. N=150. Accuracy rate is percent less than 5º absolute error. Reprinted from [54]. © 2018 IEEE.

Method	Accuracy rate	MSE (degrees^2)
1. Single model in domestic contexts	33%	60.0
2. GCC-PHAT in domestic contexts	53%	64.0
3. Combined models in domestic contexts	76%	22.1

10.3 Number of models

10.3.1 Redundant models

This work was published as Section V-C in [54]. The number of models elsewhere in the thesis, as in MOSAIC, is the same as the number of contexts. For the robot operating in the field, this means it would possess one model per environment it is expected to operate in. In this thesis, the models are pre-trained each in their own context.

The number of models used throughout the thesis has been limited to 3 (although 4 were also used in Chapter 9). There is however, a question as to how the system will perform if the number of models is increased, such that there are more models present than the number of contexts experienced. In a real-world situation it could be that a robot would build up a number of models as it experiences new contexts that cannot be dealt with by the system without generating new models.

The experiments were run with the same number of contexts (and with corresponding models trained in those contexts) as before (3 contexts), but with 4 additional models trained in other contexts. The responsibility values are shown in Figure 61. It can be seen that the models trained in a particular context do show larger responsibility values in the main, but that the relative magnitudes are now diminished as they are shared among more models. One might expect the additional models, which have not been trained in any of the contexts, to show either zero or very low responsibility values throughout the trials. Indeed, this is largely the case, but they do display some significant responsibility values in some contexts. Due to the nature of the distortions introduced by the acoustics of the environment, the responsibility is quite variable, so that models that were not trained in a particular environment can display smaller errors than those that are. It can be observed, however, that the performance is still better than that of a single model trained in all contexts. The experiments described in section 7.3.1.2 were repeated in the domestic contexts described above. In the presence of redundant models, the accuracy rate was 92%, the same as that without redundant models (i.e. that found in section 7.3.1.2), and the MSE was 5.6 degrees^2, compared to 15.8 degrees^2 for the single model. Introducing the additional models, whilst diminishing the share of the responsibility of the models that we would expect to be responsible on the basis that they were the ones that had learned in those contexts, does not appear to greatly alter the performance of the system.

10.3.2 Performance as a function of number of models and contexts

An experiment was conducted to investigate the variation in performance with the number of models. This is subtly different, but related to, the problem investigated in Section 10.3.1. In this case, the number of models matched the number of contexts experienced, as elsewhere in the thesis, that is, there were no redundant models as described in Section 10.3.1. A number of runs were conducted in which the range of contexts (value of ϕ, the angle of the sound source on its vertical axis) was varied. The reason for this is that as the number of models/contexts is varied, the difference in the value of ϕ between contexts varies too, which will affect the performance.

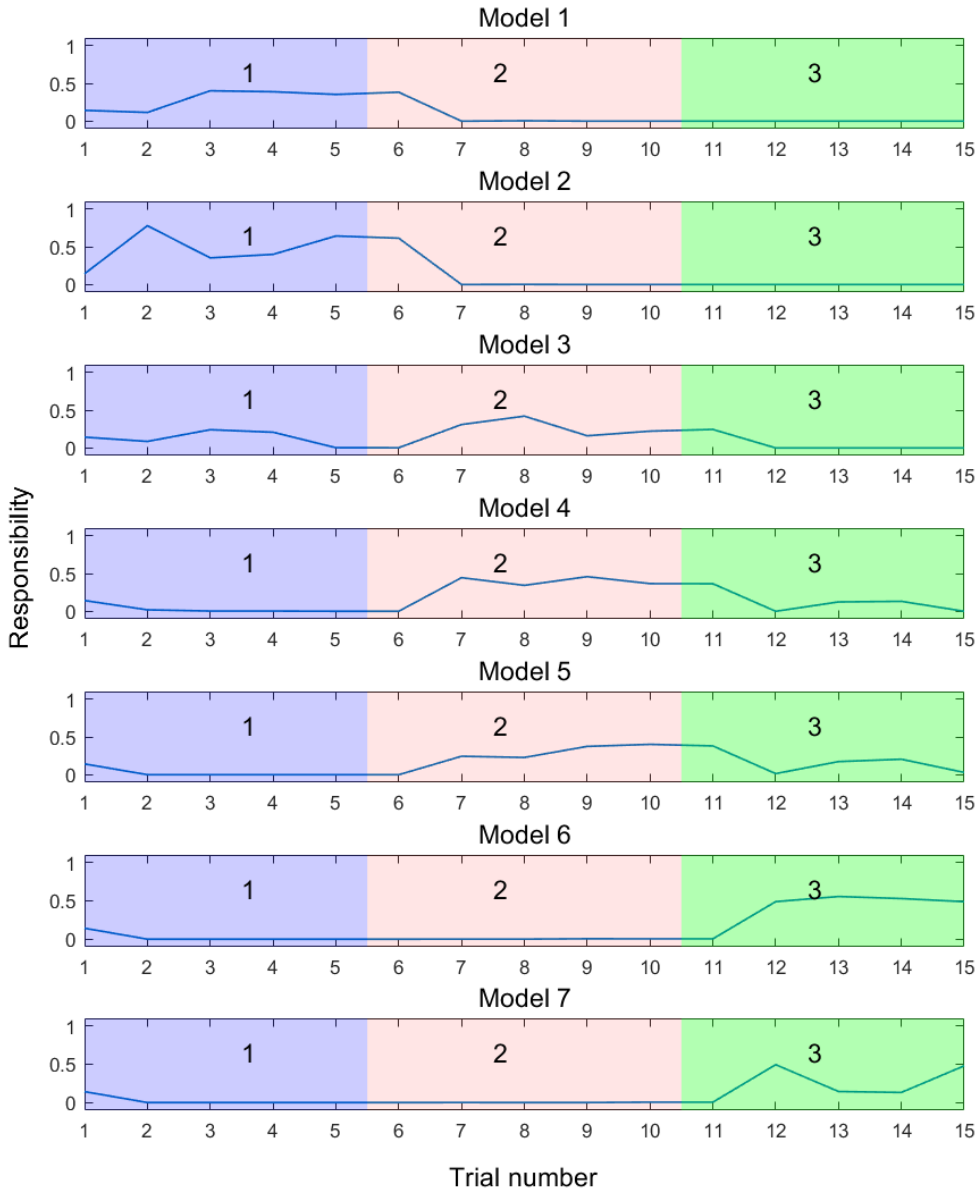


Figure 61. Responsibility signals where redundant models are present.

For each run, a range of angles were set symmetrically about 0°, with the range divided into equal increments. For example, with a range from ϕ=-90° to ϕ=90°; for two models/contexts these extremes correspond to the two contexts; for three contexts the values were ϕ=-90°, ϕ=0°, ϕ=90° whereas for four models ϕ=-90°, ϕ=-45°, ϕ=45°, ϕ=90°, for five models ϕ=-90°, ϕ=-45°, ϕ=0°, ϕ=45°, ϕ=90° and so on. This range represents the

widest used, and a variety of narrower ranges were tested. The number of models/contexts was varied between two and ten. Each model was trained in its corresponding context.

Figure 62 shows that for a narrow range in contexts (e.g. ϕ=-36^o to ϕ=36^o), there is little variation in performance with the number of models. For wider ranges, where there is greater distinction between contexts, the performance increases rapidly with increasing models/contexts for a small number of models followed by a gentler increase in performance as the number of models increases further. Also, the performance overall seems to decrease with distinctiveness between contexts. It needs to be borne in mind that this performance is measured over all contexts, not a restricted number as in Section 10.3.1, where only a subset of models had trained in the contexts presented. Figure 63 confirms similar results post de-novo learning (this also shows that post de-novo performance is somewhat worse than that of models that have been manually trained).

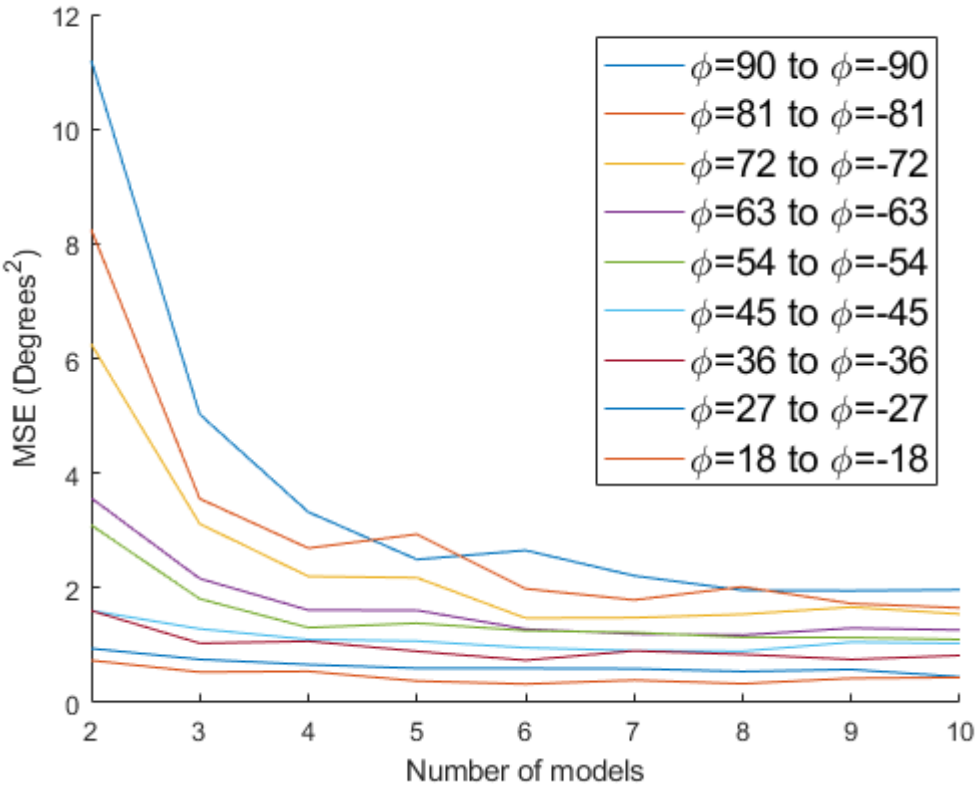


Figure 62. Performance versus number of models/contexts. Parameter is the range of contexts.

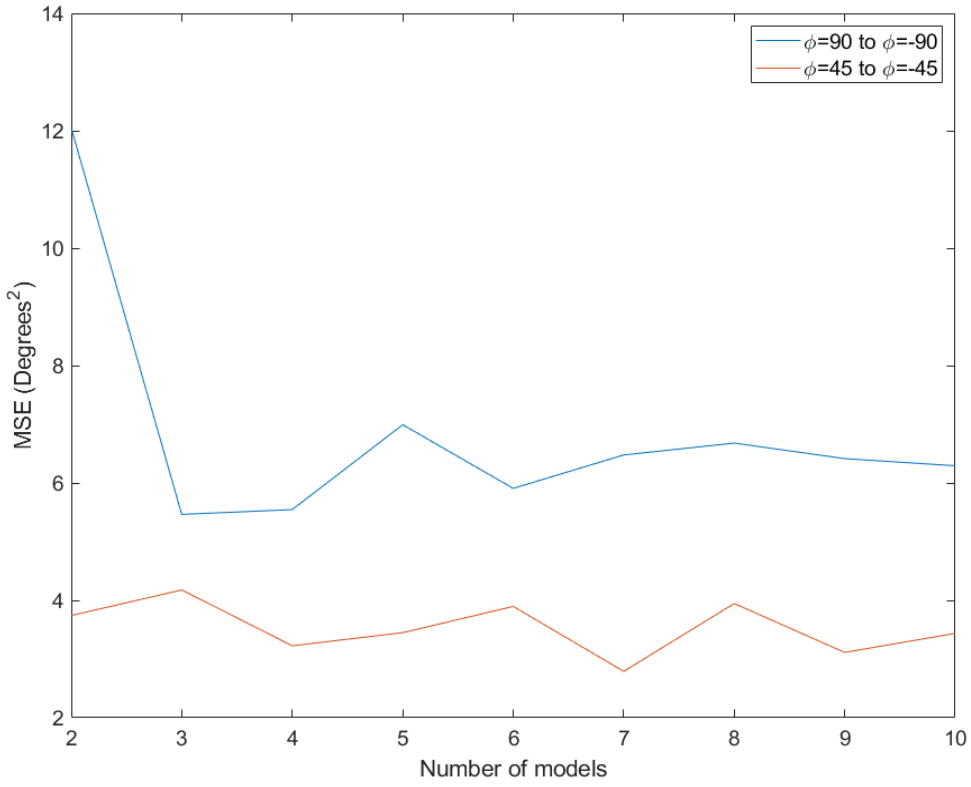


Figure 63. Performance versus number of models/contexts: de-novo models. Context range is ϕ=-90^o to ϕ=90^o (blue curve) and ϕ=-45^o to ϕ=45^o (orange curve).

10.4 Unavailability of the ground truth

The MOSAIC framework relies on ground truth becoming available, through sensory feedback, in order for the posterior responsibility values to be calculated, after a prediction has been made. However, the MOSAIC literature does not address the problem of how this would happen should the sensory feedback be disrupted for some reason, so that the ground truth becomes unavailable, so that it is not clear exactly how the system should function in the absence of the ground truth. A disruption of sensory feedback is quite possible in a real-world situation, especially a challenging one such as a disaster situation, and this appears to be an important omission. It was initially decided to use the most recently known ground truth (section 10.4.1) on the assumption that the ground truth would not change very much. However, on investigation it became clear that this was an unrealistic assumption and so investigation the mitigation of the missing ground truth using the responsibility predictor alone was then investigated (section 10.4.2).

10.4.1 Performance with ground truth missing in one trial

The work in this section was published as Section V-E in [54].

The experiment described in Section 7.3.1 was repeated but with the ground truth missing in one trial (trial 6). As mentioned above, it is not clear what should happen in this situation, but here, the last known ground truth was used. The ground truth was made unavailable during one trial (trial 6). Of course if the position of the sound source changing during the trial in which the ground truth is missing, the assumed value could be quite different from the actual value. Figure 64 shows plots of the responsibility signals of each model, with blue curves representing the output of the RE without RP involvement. In this case the dominance of model 1 is extended into context 2, where dominance of model 2 would have been expected, since that was the context in which model 2 had learned, but since the last known ground truth is from the previous context, this continued dominance by model 1 will increase the performance error. The ground truth was made available again in trial 7, and the system adjusts the responsibilities accordingly. The performance deteriorated slightly compared to the experiment where the ground truth always became available. The accuracy rate was 91% and MSE 6.4 degrees^2. However, it is reasonable to assume that prolonged absence of the ground truth will lead to further deterioration in performance. It is assumed that the RP would play a key role in mitigating this problem and this was investigated in Section 10.4.2.

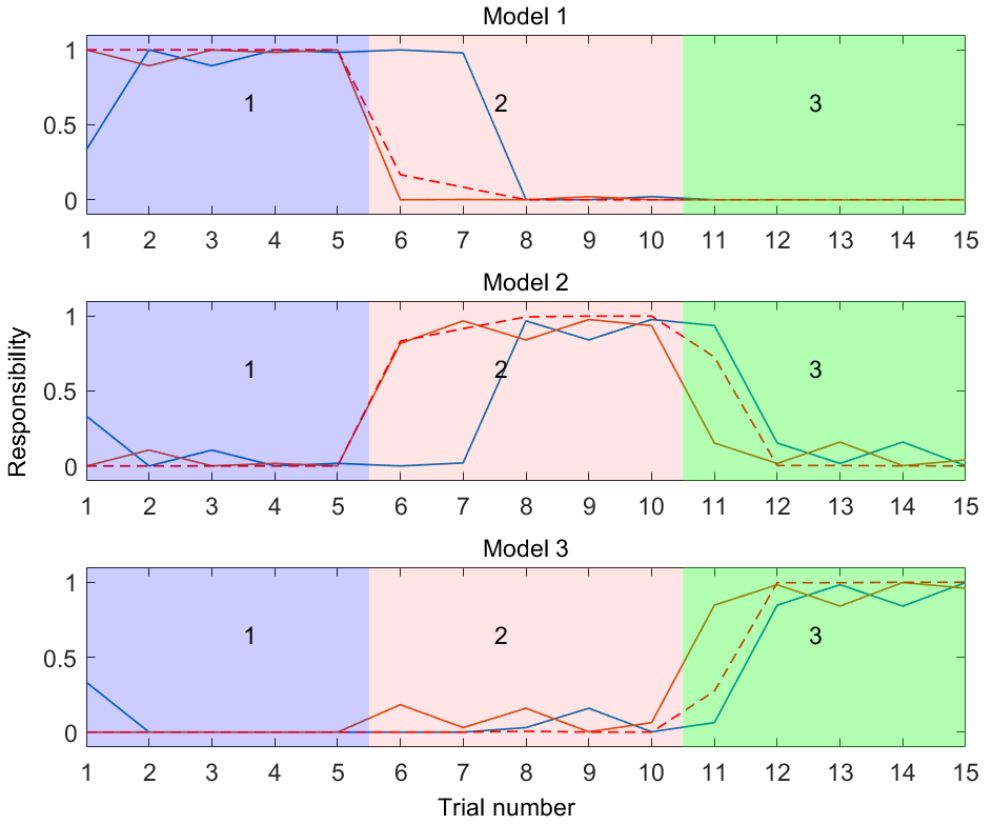


Figure 64. Responsibility signals with unavailable ground truth. In each trial the system is presented with stimulus of various azimuths in three different contexts, indicated by the coloured regions, labelled with the context number. Context 1 (blue region) is ϕ=90º left; context 2 (red region) is ϕ =0º; context 3 (green region) is ϕ=90º right. Reprinted from [54]. © 2018 IEEE.

10.4.2 Performance with ground truth missing in the presence of an RP

The work in this section was submitted to *Biomimetic and Biohybrid Systems: 8th International Conference, Living Machines 2019.*

The performance of the system was tested with an RP present and with ground truth missing from trial 6 and for the remainder of the experiment (so the ground truth is available until roughly half way through the experiment and then remaining unavailable), to investigate how the system would perform with prolonged absence of the ground truth. This seems a more probable scenario than that used in Section 10.4.1 in challenging real-world environments, where vision, for example, could become obscured through some obstruction, or the vision sensor could even become damaged. Two approaches were adopted here. First, the most recent available ground truth was used to compute the likelihood values. Second, The RP was used on its own to provide the responsibility

signals using contextual signals only (so that likelihood values are ignored while ground truth is unavailable).

Figure 65 shows responsibility signals using the first approach, where the most recently available ground truth is used to compute likelihood values. This is the same approach as adopted in Section 10.4.1. The likelihood values (reflected in the blue curve) fluctuate markedly, because the presented azimuth positions are randomly selected, whereas the most recently available ground truth will reflect the azimuth position presented in the trial in which it was determined. It should be borne in mind that a randomly fluctuating sound source position is unlikely to be encountered in a real world situation (unless, of course, the robot is moving in a random fashion). However, even if the sound source location is only slowly moving (even if it is stationary, if the robot moves, the sound source position will change with respect to the robot head), this assumed value of the ground truth could rapidly diverge form the actual value, meaning that this may not be a practical approach to take. While the RP output (solid orange curve) more closely follows the responsibility pattern we would have expected were the ground truth available, the overall responsibility (red broken curve) is strongly influenced by the likelihood values. The accuracy rate using this approach was 83% and MSE of 14.5 degrees^2. This compares to 71% and 19.9 degrees^2 respectively without the RP (using the same data set), so there is a modest improvement in performance.

Figure 66 shows the responsibility values where the RP alone provides the responsibility values based on the extracted audio feature. In this case the overall responsibility (red broken curve) now exactly follows the RP output, as expected. Using this approach, the accuracy rate was 99% and MSE was 1.7 degrees^2, a marked improvement compared to that without RP (and where the most recent available ground truth was used) and is comparable to the performance of the system in the presence of an RP where the ground truth is available throughout the experiment (accuracy rate 100% and MSE 1.12 degrees^2 using the same data set).

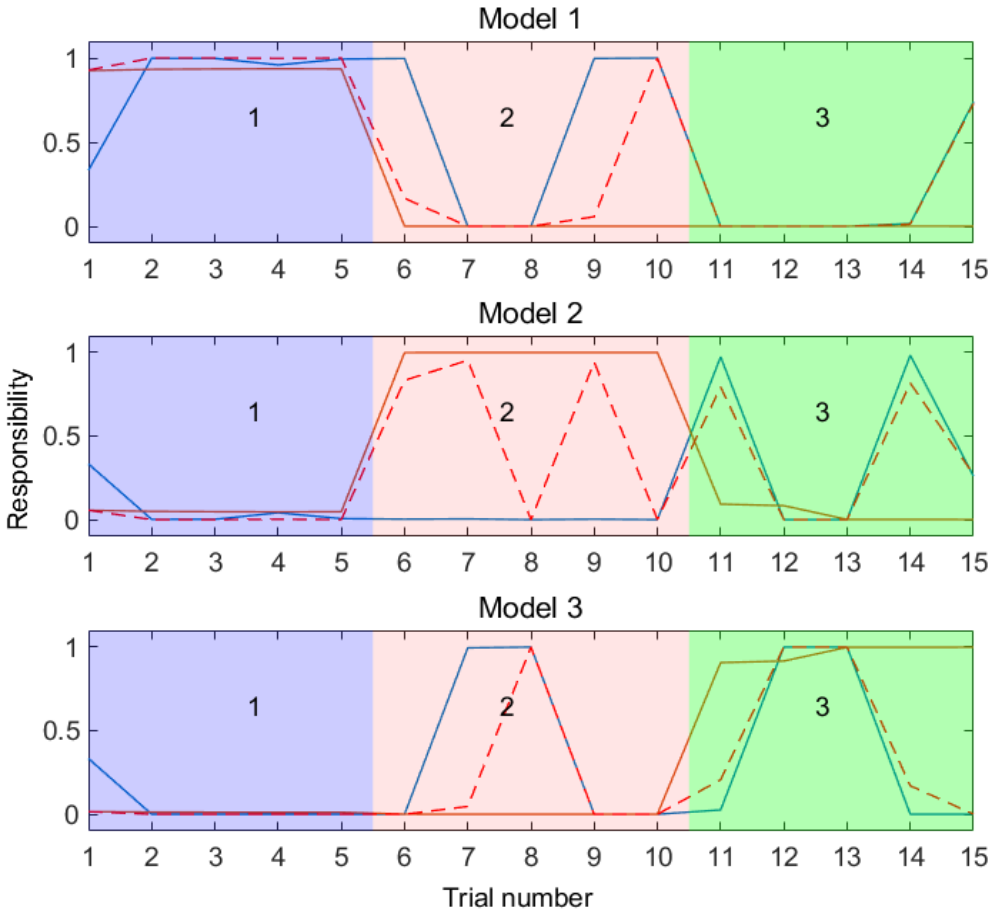


Figure 65. Responsibility signals with unavailable ground truth: using most recent. Ground truth becomes unavailable from trial 6, and the most recently available value is used to calculate the responsibility. In each trial the system is presented with stimulus of various azimuths in three different contexts, indicated by the coloured regions, labelled with the context number. Context 1 (blue region) is ϕ=90^o left; context 2 (red region) is ϕ =0^o; context 3 (green region) is ϕ=90^o right.

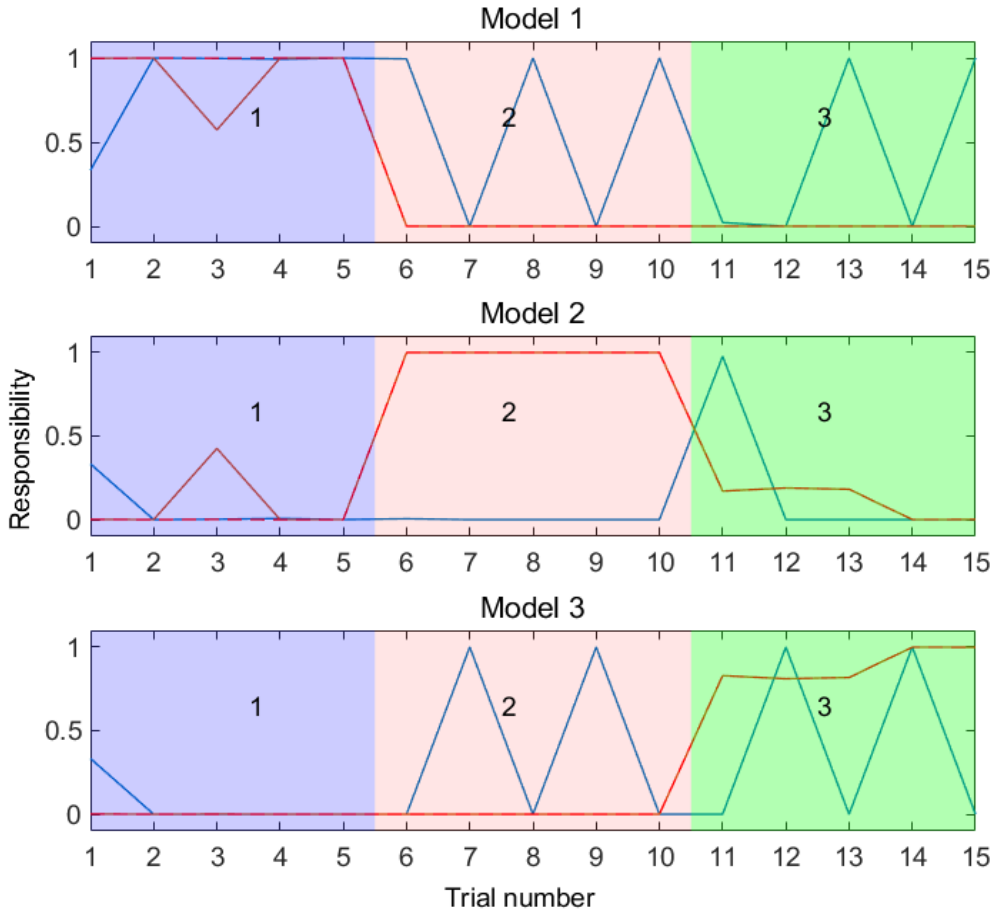


Figure 66. Responsibility signals with unavailable ground truth: RP only. Ground truth becomes unavailable from trial 6. The RP only is used to provide responsibility signals when the ground truth becomes unavailable. In each trial the system is presented with stimulus of various azimuths in three different contexts, indicated by the coloured regions, labelled with the context number. Context 1 (blue region) is ϕ=90º left; context 2 (red region) is ϕ =0º; context 3 (green region) is ϕ=90º right.

10.5 Improving robustness: using other SSL algorithms

As explained in Section 1.1 the system sits on top of an SSL algorithm, which in this case was a quite basic algorithm that is not robust to background noise, reverberation, multiple sources and so on. It ought to be possible to "plug-in" a different SSL algorithm, and if the system performs as expected from this thesis, it should learn/adapt with the new algorithm. In this sense, the system relies on the robustness of the underlying SSL

algorithm and the potential of the system calibrate what could be a sophisticated, robust SSL algorithm in multiple unstructured environments.

The basic cross-correlation algorithm used in previous chapters was replaced by the GCC-PHAT SSL algorithm. GCC-PHAT is a popular SSL algorithm as it is considered to be robust to reverberation [145]. Models were retrained (3 models, 3 contexts) using GCC-PHAT. The accuracy rate of the combined models post-learning improved from 92% with the basic cross-correlation algorithm to 95% with GCC-PHAT as the SSL algorithm. The MSE improved slightly from 5.8 degrees^2 with basic cross-correlation to 5.6 degrees^2 with GCC-PHAT. Clearly, this is not a great improvement in performance, however, it does demonstrate the flexibility of the proposed system in that a different SSL algorithm can be successfully substituted. It would be interesting to see how the system with GCC-PHAT would perform in more challenging environments such as those investigated in Section 10.2, or indeed, how this might improve with an even more sophisticated SSL algorithm, including multiple microphone arrays, active SSL and so on.

10.6 Chapter summary

This chapter demonstrates the weakness of the basic system insofar as it is based on the original MOSAIC framework and a simple SSL algorithm. Nevertheless, the performance of the system is respectable and in some situations (such as with missing ground truth, with the RP alone providing responsibility signals) quite impressive, and the chapter demonstrates that even with a quite unsophisticated underlying SSL algorithm, the system improves on the performance of that SSL alone. The system was taken out of the experimental arena, and still performed relatively well in a more challenging domestic environment (although the conditions were still quite constrained). The chapter has shown that the system performs well when the number of models (and hence, the number of experienced contexts) increases, although this was tested only up to 10 models/contexts. This held true for models that learned de-novo, although the distinctiveness with which the models divided up experience deteriorated with the number of models/contexts. The chapter also demonstrated that the underlying SSL algorithm can be substituted for a different one, which led to a slight improvement in performance. However, there are now a wealth of SSL techniques that could be used, in principle.

Chapter 11 Conclusions and future work

11.1 Conclusions

This thesis has demonstrated two key novel concepts, and 2 secondary concepts.

First, it has demonstrated that an adaptive filter model of the cerebellum can be successfully applied to the calibration of a SSL algorithm in a particular acoustic context, learning an azimuth-dependent error in SSL caused by the acoustic properties of the environment (Chapter 4). Specifically, Chapter 4 demonstrated the basic SSL calibration technique, including a pilot experiment to confirm that vision could be used to generate a teaching signal, although the odometry of the experimental apparatus was used to determine the ground truth azimuth throughout the remainder of the thesis. The calibrated SSL outperformed the un-calibrated SSL by a considerable margin (Section 4.3.2). This specifically addresses research question 1: *Is it possible to apply cerebellar calibration to SSL?* The chapter also demonstrated that the system performed well when a pure tone was used rather than the Gaussian noise used throughout the thesis; this is significant as pure tones are somewhat difficult to localise.

Second, the thesis has demonstrated that a multiple-models inspired approach (specifically, inspired by the MOSAIC framework), although developed in the context of motor control, can be successfully re-purposed, with adaptation, to select a set of calibration models for different acoustic environments (Chapter 6 and Chapter 7), and this approach appears to be completely unique in robot audition. This specifically addresses research question 2: *Is it possible to implement a multiple-models inspired architecture that will select an appropriate cerebellar calibration model, or set of models, in different acoustic contexts?* Specifically, the combined models approach outperformed the un-calibrated SSL estimates and a single model trained in all contexts (which latter is in agreement with the claim made for MOSAIC that a single general model that could operate across multiple contexts would be too complex). This is, perhaps, the most significant outcome of the thesis.

Third, the thesis has demonstrated the utility of responsibility prediction, proposed in the MOSAIC literature, in improving the robustness of a multiple models framework to the unavailability of ground truth, which is *not* covered in the MOSAIC literature, where the ground truth is assumed to always become available through sensory feedback. This

aspect of the thesis emerged at a later stage in the project, after the research questions were drawn up, yet it appears to be a particularly significant result.

Last, the utility of the so-called *cerebellar chip* has been further demonstrated through the development of a new audio application that makes use of the adaptive filter model of the cerebellum as well as through the development of an RP that is also based on the adaptive filter model of the cerebellum, rather than the more conventional function fitting NN used in the literature.

The work in Chapter 6 demonstrated the application of a multiple-models approach to identifying the acoustic environment, and the system was able to correctly identify the acoustic environment in 69.4% of the 49 cases tested (7 different azimuths in 7 different acoustic contexts- see section 6.3), which although not particularly high is considerably better than chance (around 14%). Because the system used the calibration error of each model, the performance of the system was poor where there was little difference between the different models' estimates; often this was where there was little error introduced into the SSL estimation, so that models trained in similar contexts were confused. Although there may be further work that could be done with identification of the environment this is not the main thrust of the thesis, and chapter 6 was more a demonstration that the system identifies the context based on the model that produces the *lowest error* in its estimate of the sound source location, even if a model is selected that did not learn in the given environment, leading to a misidentification of the environment. The question arises as to whether the system could select the most appropriate model (or even set of models) for calibration, so that Chapter 7 focuses on improving the SSL calibration in a particular environment by combining the outputs of all the models, in proportion to how well each model is able to calibrate the SSL estimate in a particular environment. Chapter 7 showed that combining the outputs of multiple models in this way improved the SSL estimate with a MSE of 5.4 degrees^2 compared to 15.8 degrees^2 for a single model that had been trained in all contexts. The chapter also showed that combining the outputs of models in combination with their responsibilities was also an improvement compared to *switching* to the single best model (in fact, the switched version was not much better than a general single model, with an MSE of 12 degrees^2). As such, Chapter 7 perhaps represents the most significant outcome of the thesis.

Chapter 8 demonstrated that prior prediction based on the inclusion of an RP improved the performance of the system, particularly during transitions between acoustic contexts, where the context has changed but the responsibility values have not yet updated. This is in line with the MOSAIC literature, but the thesis goes further, to address the problem of the ground truth through sensory feedback becoming unavailable, which is a realistic prospect in challenging real world environments such as the aftermath of a disaster. This latter result is significant as it suggests a way forward in dealing with unreliable sensor input in challenging environments. The chapter also showed, in line with the MOSAIC literature, that a misclassification of the acoustic context by the RP is corrected *a-posteriori* by the RE when the ground truth becomes available. However, there is still a question of how this would work if sensory feedback is disrupted in a challenging environment and the ground truth does not become available.

The system has been based on the original MOSAIC framework, and it could be that more sophisticated approaches, such as HMM-MOSAIC, could point towards a more practical system. Nonetheless, this use of multiple models, and specifically an adapted version of MOSAIC, which was developed in the context of motor control, has not previously been used for audio applications, making this approach unique.

Addressing research question 3: *Can a multiple-models inspired audio calibration system self-organise?* Chapter 9 showed that the proposed system has the potential to self-organise, from a de-novo state, that is, with a pre-determined number of models (hence, although this follows the approach taken in the MOSAIC literature, it is not true self organisation). This was a somewhat secondary research question, which the author added out of curiosity, and since it was also a claim of the MOSAIC framework. The distinctiveness with which the models divide up the experience appears to decrease with the number of models, although, as with the manually trained models, the SSL calibration performance does not decrease with the number of models. Chapter 9 also demonstrated a "complete" system as it were, with RPs learning alongside (to an extent) the calibration models. This worked particularly well were the RPs learned post-de-novo learning of the models, as was the case in the MOSAIC literature. An attempt to move somewhat towards a situation in which the RPs learned in batches "alongside" the models proved much more difficult. It may be that a more sophisticated approach to training the RPs may be required (for example, there is no stopping criterion, with the training iterations set through trial

and error), or a more sophisticated version of the MOSAIC framework as discussed above.

Chapter 10 made a preliminary investigation of a number of issues that might occur as the system moves towards being used in real-world environments. Key findings were that the system performed respectably in a more challenging environment away from the experimental arena (albeit with a reduced performance), considerably outperforming a single model; that the system was able to cope with missing ground-truth, especially with the use of the RP providing responsibility signals where the absence of ground truth is prolonged- although this depended on RPs that had been pre-trained; SSL calibration performance appeared largely independent of the number of models and contexts used (up to 10 models/contexts were tested); it was possible to substitute a different SSL algorithm for the one used in the remainder of the thesis. Although these experiments were somewhat preliminary and limited in scope, the chapter does indicate that it is worth investigating the real world performance of the system, perhaps with more sophisticated approaches (e.g. to SSL and multiple models selection) than those investigated in this thesis. The chapter highlighted the limitations of the proposed system in more challenging environments, but that it was still able to perform, and the indication is that a more sophisticated system could be developed with greater robustness.

11.2 Limitations of the work

There are two key limitations of the thesis and the methods employed. First, as mentioned in section 4.1.1 the cerebellar calibration model lacks a stopping criteria during learning and the literature that describes the model does not offer a rationale for choice of learning trial number. This is also true for the RP as mentioned in section 8.7 and both the calibration model and the RP implementation is somewhat crude when compared to more well-established neural networks, for example, lacking a stopping criteria during training. Other parameters could be investigated such as tuning the learning rate and investigating different configurations of parallel fibre and basis filters- this is especially true of the RP model which it is suspected is far from optimal in its implementation, being a proof-of-concept model. In this thesis, the cerebellar models were seen as proof of concept, and future work will need to develop these models into more practical implementations.

Second, as mentioned section 9.4, the thesis has adopted the same approach to de-novo learning as described in Haruno et al. [122], which is not true self organisation. As such

research question 3 *can a multiple-models inspired audio calibration system self-organise?* has only been partially answered. This can be justified because as mentioned in section 11.1, this was a somewhat secondary research question which was formulated later in the project, and to fully answer this question would have distracted from answering the two main research questions. A fuller answer to that research question is left for future work.

11.3 Future work

This thesis has posed a number of questions that still need to be answered, such as, how will such a system create new models and under what circumstances (i.e. how would it truly self-organise); how would the system perform under real-world conditions, such as multiple sound sources, background noise, reverberation and so on; is the system transferable to a mobile platform with the concomitant lack of resources and power, along with the need to operate in real-time; how would calibration models and RPs learn in a real-world scenario (as mentioned in section 11.2, the thesis has assumed that the robot is able to pre-train in "ideal" conditions). The thesis has touched on sensor fusion and has highlighted the potential for this in future work.

11.3.1 Development of a practical framework

As mentioned in Chapter 10 and in section 11.2, the system proposed in this thesis has been tested under constrained conditions and using cerebellar models that are quite crude in implementation. Also, there are a number of questions about how a practical system would operate, especially about how the system would self-organise and evolve in a real world situation. As mentioned in Section 3.3, there is a potential tie-in to the subsumption architecture and this might be a fruitful approach to investigate. HMM-MOSAIC may be a potential area for investigation, especially in light of the claim that it allows automatic computation of the responsibility scaling factor σ.

11.3.2 Implementation on other platforms

11.3.2.1 Hardware implementations

As mentioned in Section 11.3, real-world implementations of the proposed system are likely to require real-time operation with low power consumption. As mentioned in Section 4.2 a Zynq SoC was used to implement a gammatone filter bank at an early stage (although this was not used) and this type of device, that allows an arbitrary design split between hardware and software on the same fabric could be a candidate for compact, real-

time operation. Neuromorphic engineering, mentioned in Section 2.2.1, provides a low power consumption, spike based potential solution, and it would be an interesting question as to how the multiple models calibration system could be transferred to such a platform.

11.3.2.2 Mobile platforms

Implementation of the system on a mobile platform was eschewed in the thesis, although the system was developed with eventual migration to a mobile platform in mind throughout the thesis. This would be particularly useful in investigating sensor fusion approaches using this system in more realistic situations, and how the system might perform as the robot moves toward the target sound source. In the thesis, the performance of the system was measured on a stationary platform, with the sound source presented in many experiments from random directions. It is unlikely in a real world scenario that the sequence of sound source azimuths would evolve in such a way, and a mobile platform would be the only realistic way of testing such a scenario.

11.3.3 Extension to 2 or 3 dimensional SSL

As mentioned in Section 2.3 it ought to be possible to extend the SSL system to 2 dimensions, as a 2D map has already been successfully calibrated in precursory work [4]. If this were to include extension to estimation of elevation, it would require the investigation of other binaural cues such as ILD and techniques such as asymmetrical pinnae. This does lead, however, to further questions such as will the cerebellar calibration system be able to compensate for manufacturing imperfections in, or damage to the artificial pinnae for example. In fact, cerebellar calibration ought to be well suited to do just this. Full 3- dimensional SSL would include an estimation of distance to source. This is somewhat more problematic than azimuth or elevation estimation, and techniques are not well established, although they have been investigated for binaural systems [146], and this could be a particularly interesting area of investigation for application of such a system.

11.3.4 Unavailability of ground truth

The problem of the ground truth becoming unavailable was discussed in Section 10.4. This issue is not discussed in the MOSAIC literature; it is assumed that sensory feedback will always become available, and it is not clear how a robot would deal with this situation. In this thesis, two approaches were investigated- the last known ground truth being used and falling back on the RP only when ground truth becomes unavailable. The

approach using last known ground truth provided a reasonable performance where the ground truth was only missing for one trial (Section 10.4.1), but was poor with prolonged absence of the ground truth (Section 10.4.2). This is perhaps to be expected as the ground truth could differ markedly from the assumed value as the robot navigates its environment, so this would not seem to be a particularly practical approach. Performance where the responsibilities were provided by the RP alone was impressive with prolonged absence of the ground truth, although this does rely on the RP having previously learned against its model. That is to say, ground truth would need to be available during the learning of the RPs. Approaches to the unavailability of ground truth seems a fruitful area for further investigation, especially as it is not covered in the literature on MOSAIC, and may be linked to work on sensor fusion.

11.3.5 Application of cerebellar calibration to other areas of Robot Audition

Other areas that the approaches covered in this thesis could be applied to are:

- ego-noise cancellation in multiple contexts
- emotion recognition (especially recognising multiple types of emotion or those of different people)
- sound source recognition
- non-binaural and active SSL

The adaptive filter model of the cerebellum ought to be particularly well suited to ego-noise cancellation. Ego-noise cancellation in itself is not new, but the use of multiple models to adapt to different contexts shows promise where various aspects of ego noise could be adapted to (as the robot could generate acoustic noise in a variety of ways in different situations). With emotion and sound source recognition, for example, one could imagine a system which discriminates between different sound sources and attends to the one of most interest. Possible questions here could include whether different sound sources would represent different contexts, and whether the system would focus on a sound source based on responsibility signals, or whether the discrimination would be carried out by a subsystem. The thesis has eschewed active SSL in favour of passive, binaural SSL. The basic system used in the thesis is certainly not robust and it could be worth investigating the proposed system with more up-to-date SSL algorithms, including active SSL in which the robot itself emits sounds or orients its sensors (much in the same way as an animal might re-orient its head to aid in the localisation of sound).

11.3.6 Improving robustness in real-world situations

A small step in this direction has been taken by substituting a different SSL algorithm for the one used. It would be interesting to see to what extent the system enhances the performance of a more robust and sophisticated SSL algorithm in real, challenging situations (use of the system developed in this thesis with a basic SSL algorithm in domestic contexts resulted in a quite poor performance, although the system performed considerably better than a single model).

11.3.7 Sensor fusion

This thesis has touched on sensor fusion, but only to a limited extent in the sense that vision was considered as a potential means of providing sensory feedback about the ground truth position of the sound source. This was confirmed as a valid approach in a pilot experiment described in Section 4.2.2, but no further investigation was carried out. The proposed system described in the thesis offers the potential for a future focus on sensor fusion, and indeed this could go beyond audio-visual integration, and include other senses such as touch and olfaction, both of which would potentially be powerful candidates for sensory input in a disaster situation or in challenging and extreme environments, either in providing the ground truth for the RE, or in providing contextual signals for the RP. This fusion could be with other sensory modalities providing contextual signals for the RP, or alternative localisation schemes. One could imagine "cerebellar chips" with a rich set of inputs (inspired by the richness of inputs to the real cerebellum) with many types of inputs of differing modalities, adapting to many different, challenging situations.

11.3.8 Self-organisation

It has already been discussed (Section 9.1) that the cerebellar models can learn from a de-novo state (with a fixed number of pre-initialised models). Although this replicates what was done within MOSAIC, this is not true self-organisation, as the models do not emerge from a tabula rasa state. A robot operating in a real world scenario ought to be able to learn from experience rather than by pre-ordained design, and further work will be required in this area, perhaps including SOMs as a potential basis for investigation of this, although there may of course be other approaches.

Chapter 12 References

[1] E. Garcia-Fidalgo and A. Ortiz, "Vision-based topological mapping and localization methods: A survey," *Robotics and Autonomous Systems,* vol. 64, pp. 1-20, 2015.

[2] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech & Language,* vol. 34, pp. 87-112, 2015.

[3] (2015). *Bioinspired Control of Electro-Active Polymers for Next Generation Soft Robots*. Available: https://gtr.ukri.org/projects?ref=EP%2FI032533%2F1

[4] 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)*, 2016, pp. 967-972.

[5] E. D. Wilson, S. R. Anderson, P. Dean, and J. Porrill, "Sensorimotor maps can be dynamically calibrated using an adaptive-filter model of the cerebellum," *PLOS Computational Biology,* vol. 15, p. e1007187, 2019.

[6] D. M. Wolpert and M. Kawato, "Multiple paired forward and inverse models for motor control," *Neural Networks,* vol. 11, pp. 1317-1329, 1998.

[7] 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, pp. 789-796, 2004.

[8] J. Porrill, P. Dean, and S. R. Anderson, "Adaptive filters and internal models: Multilevel description of cerebellar function," *Neural Networks,* vol. 47, pp. 134-149, 2013.

[9] D. Bodznick and J. Montgomery, *Evolution of the Cerebellar Sense of Self*. Oxford: Oxford University Press, 2016.

[10] J. Schnupp, *Auditory Neuroscience: Making Sense of Sound*: MIT Press, 2012.

[11] M. Slaney. (1998, 1/7/2015). *Auditory Toolbox*. Available: http://rvl4.ecn.purdue.edu/~malcolm/interval/1998-010/

[12] P. L. Søndergaard and P. Majdak, "The Auditory Modeling Toolbox," in *The Technology of Binaural Listening*, J. Blauert, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 33-56.

[13] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, 1982, pp. 1282-1285.

[14] R. D. Patterson, et al., "Complex sounds and auditory images," *Auditory physiology and perception,* vol. 83, pp. 429-446, 1992.

[15] D. Sabo, S. Weiss, and M. Furst, "A Parallel Algorithm for a Physiological Non-linear Model of the Cochlea," *Procedia Computer Science,* vol. 18, pp. 682-691, 2013.

[16] C. Binbin, Y. Fei, and X. Jing, "A biomimetic model for bats' echolocation signal processing and it's implementation on FPGA," in *Signal Processing Systems (ICSPS), 2010 2nd International Conference on*, 2010, pp. V1-452-V1-456.

[17] I. Gambin, I. Grech, O. Casha, E. Gatt, and J. Micallef, "Digital cochlea model implementation using Xilinx XC3S500E Spartan-3E FPGA," in *Electronics, Circuits, and Systems (ICECS), 2010 17th IEEE International Conference on*, 2010, pp. 946-949.

[18] R. F. Lyon and C. Mead, "An analog electronic cochlea," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 36, pp. 1119-1134, 1988.
[19] V. Y. Chan, C. T. Jin, and A. van Schaik, "Adaptive sound localization with a silicon cochlea pair," *Frontiers in Neuroscience,* vol. 4, p. 196, 2010.
[20] V. Y. Chan, C. T. Jin, and A. van Schaik, "Neuromorphic audio-visual sensor fusion on a sound-localizing robot," *Front Neurosci,* vol. 6, p. 21, 2012.
[21] A. G. Katsiamis, E. Drakakis, and R. F. Lyon, "A biomimetic, 4.5 w, 120+ dB, log-domain cochlea channel with AGC," *Solid-State Circuits, IEEE Journal of,* vol. 44, pp. 1006-1022, 2009.
[22] C. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE,* vol. 78, pp. 1629-1636, 1990.
[23] F. Galluppi, X. Lagorce, E. Stromatias, M. Pfeiffer, L. A. Plana, S. B. Furber*, et al.*, "A framework for plasticity implementation on the SpiNNaker neural architecture," *Frontiers in neuroscience,* vol. 8, p. 429, 2014.
[24] V. Chan, "Audio-visual sensor fusion for object localization," 2009.
[25] V. Y.-S. Chan, C. T. Jin, and A. van Schaik, "Neuromorphic audio-visual sensor fusion on a sound-localizing robot," *Frontiers in neuroscience,* vol. 6, p. 21, 2012.
[26] H. Finger and L. Shih-Chii, "Estimating the location of a sound source with a spike-timing localization algorithm," in *Circuits and Systems (ISCAS), 2011 IEEE International Symposium on*, 2011, pp. 2461-2464.
[27] A. Katsiamis, E. Drakakis, and R. Lyon, "Practical Gammatone-Like Filters for Auditory Processing," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 2007, p. 063685, 2007.
[28] J. Huo, A. Murray, and D. Wei, "Adaptive Visual and Auditory Map Alignment in Barn Owl Superior Colliculus and Its Neuromorphic Implementation," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 23, pp. 1486-1497, 2012.
[29] T. L. K. Nakadai, H. G. Okuno and H. Kitano, "Active audition for humanoid," in *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 2000.
[30] H. G. Okuno, T. Ogata, K. Komatani, and K. Nakadai, "Computational auditory scene analysis and its application to robot audition," in *International Conference on Informatics Research for Development of Knowledge Society Infrastructure, 2004. ICKS 2004.*, 2004, pp. 73-80.
[31] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5610-5614.
[32] S. Argentieri, A. Portello, M. Bernard, P. Danès, and B. Gas, "Binaural Systems in Robotics," in *The Technology of Binaural Listening*, J. Blauert, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 225-253.
[33] A. Badali, J. M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 2033-2038.
[34] X. Li, M. Shen, W. Wang, and H. Liu, "Real-time Sound Source Localization for a Mobile Robot Based on the Guided Spectral-Temporal Position Method," *International Journal of Advanced Robotic Systems,* vol. 9, 2012.

[35] S. Kagami, S. Thompson, Y. Sasaki, H. Mizoguchi, and T. Enomoto, "2D sound source mapping from mobile robot using beamforming and particle filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3689-3692.

[36] G. Narang, K. Nakamura, and K. Nakadai, "Auditory-aware navigation for mobile robots based on reflection-robust sound source localization and visual SLAM," in *IEEE International Conference on Systems, Man and Cybernetics*, 2014, pp. 4021-4026.

[37] M. Rucci, J. Wray, and G. M. Edelman, "Robust localization of auditory and visual targets in a robotic barn owl," *Robotics and Autonomous Systems,* vol. 30, pp. 181-193, 2000.

[38] H. G. Okuno, K. Nakadai, T. Lourens, and H. Kitano, "Sound and Visual Tracking for Humanoid Robot," *Applied Intelligence,* vol. 20, pp. 253-266, May 01 2004.

[39] E. Martinson and A. Schultz, "Discovery of sound sources by an autonomous mobile robot," *Autonomous Robots,* vol. 27, pp. 221-237, 2009.

[40] C. Youngkyu, D. Yook, C. Sukmoon, and K. Hyunsoo, "Sound source localization for robot auditory systems," *Consumer Electronics, IEEE Transactions on,* vol. 55, pp. 1663-1668, 2009.

[41] H. Finger, S. C. Liu, P. Ruvolo, and J. R. Movellan, "Approaches and databases for online calibration of binaural sound localization for robotic heads," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4340-4345.

[42] C. Evers and P. Naylor, "Acoustic SLAM," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 26, pp. 1484-1498, 2018.

[43] J. Blauert, *Spatial hearing: the psychophysics of human sound localization* vol. Revised. Cambridge, Mass;London;: MIT Press, 1997.

[44] M. Rucci, "Robust localization of auditory and visual targets in a robotic barn owl," *Robotics and Autonomous Systems,* vol. 30, pp. 181-193, 2000.

[45] B. Glackin, J. A. Wall, T. M. McGinnity, L. P. Maguire, and L. J. McDaid, "A spiking neural network model of the medial superior olive using spike timing dependent plasticity for sound localization," *Front Comput Neurosci,* vol. 4, 2010.

[46] J. Liu, D. Perez-Gonzalez, A. Rees, H. Erwin, and S. Wermter, "A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation," *Neurocomputing,* vol. 74, pp. 129-139, 2010.

[47] S. Hwang, Y. Park, and Y.-s. Park, "Sound direction estimation using an artificial ear for robots," *Robotics and Autonomous Systems,* vol. 59, pp. 208-217, 2011.

[48] A. Umbarkar, V. Subramanian, and A. Doboli, "Low-cost sound-based localization using programmable mixed-signal systems-on-chip," *Microelectronics Journal,* vol. 42, pp. 382-395, 2011.

[49] V. S. Hanson and K. M. Odame, "Real-time source separation on a field programmable gate array platform," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012, pp. 2925-2928.

[50] L. Zhang, J. Hallam, J. Christensen-dalsgaard, and J. Christensen-Dalsgaard, "Effects of asymmetry and learning on phonotaxis in a robot based on the lizard auditory system," *Adaptive Behavior,* vol. 20, pp. 159-171, 2012.

[51] S. Lee, Y. Park, and J.-S. Choi, "Estimation of multiple sound source directions using artificial robot ears," *Applied Acoustics,* vol. 77, pp. 49-58, 2014.
[52] S. Mattes, P. Nelson, F. Fazi, and M. Capp, "Exploration of a Biologically Inspired Model for Sound Source Localization In 3D Space," *Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, 2014,* 2014.
[53] L. A. Jeffress, "A place theory of sound localization," *Journal of Comparative and Physiological Psychology,* vol. 41, pp. 35-39, 1948.
[54] M. D. Baxendale, M. J. Pearson, M. Nibouche, E. L. Secco, and A. G. Pipe, "Audio Localization for Robots Using Parallel Cerebellar Models," *IEEE Robotics and Automation Letters,* vol. 3, pp. 3185-3192, 2018.
[55] A. A. Handzel, "High acuity sound-source localization by means of a triangular spherical array," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 2005, pp. iv/1057-iv/1060 Vol. 4.
[56] J. Huo and A. Murray, "The adaptation of visual and auditory integration in the barn owl superior colliculus with Spike Timing Dependent Plasticity," *Neural Networks,* vol. 22, pp. 913-921, 2009.
[57] J. A. Wall, T. M. McGinnity, and L. P. Maguire, "A comparison of sound localisation techniques using cross-correlation and spiking neural networks for mobile robotics," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 2011, pp. 1981-1987.
[58] K. Youssef, S. Argentieri, and J. L. Zarader, "A binaural sound source localization method using auditive cues and vision," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 217-220.
[59] J. Davila-Chacon, S. Magg, L. Jindong, and S. Wermter, "Neural and statistical processing of spatial cues for sound source localisation," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, 2013, pp. 1-8.
[60] F. Keyrouz, Y. Naous, and K. Diepold, "A New Method for Binaural 3-D Localization Based on Hrtfs," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006, pp. V-341 - V-344.
[61] F. Keyrouz, "Advanced Binaural Sound Localization in 3-D for Humanoid Robots," *Instrumentation and Measurement, IEEE Transactions on,* vol. 63, pp. 2098-2107, 2014.
[62] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 24, pp. 320-327, 1976.
[63] J. Kelly and G. S. Sukhatme, "Visual-inertial sensor fusion: localization, mapping and sensor-to- sensor self-calibration.(Technical report)," *The International Journal of Robotics Research,* vol. 30, pp. 56-79, 2011.
[64] H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based online calibration of asynchronous microphone array for robot audition," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 524-529.
[65] L. Battista, E. Schena, G. Schiavone, S. A. Sciuto, and S. Silvestri, "Calibration and Uncertainty Evaluation Using Monte Carlo Method of a Simple 2D Sound Localization System," *Sensors Journal, IEEE,* vol. 13, pp. 3312-3318, 2013.

[66] S. Chachada and C. C. J. Kuo, "Environmental sound recognition: a survey," *APSIPA Transactions on Signal and Information Processing,* vol. 3, p. e14, 2014.
[67] S. Chu, S. S. Narayanan, C. C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, 2006, pp. 885-888.
[68] N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, and H. G. Okuno, "Environmental Sound Recognition for Robot Audition Using Matching-Pursuit," in *Modern Approaches in Applied Intelligence*, M. C. K. Mehrotra K.G., Oh J.C., Varshney P.K., Ali M., Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1-10.
[69] V. Bisot, R. Serizel, S. Essid, G. Richard, V. Bisot, R. Serizel*, et al.*, "Feature Learning With Matrix Factorization Applied to Acoustic Scene Classification," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.,* vol. 25, pp. 1216-1229, 2017.
[70] U. S. Prakruthi, D. Kiran, and H. Ramasangu, "High performance neural network based acoustic scene classification," in *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 781-784.
[71] M. Fujita, "Adaptive filter model of the cerebellum," *Biol Cybern,* vol. 45, pp. 195-206, 1982.
[72] P. Dean, J. Porrill, C. F. Ekerot, H. Jorntell, and C.-F. Ekerot, "The cerebellar microcircuit as an adaptive filter: experimental and computational evidence.(Report)," *Nature Reviews Neuroscience,* vol. 11, p. 30, 2010.
[73] R. Apps and M. Garwicz, "Anatomical and physiological foundations of cerebellar information processing," *Nature Reviews Neuroscience,* vol. 6, pp. 297-311, 2005.
[74] S. R. Anderson, J. Porrill, M. J. Pearson, and A. G. Pipe, "An internal model architecture for novelty detection: implications for cerebellar and collicular roles in sensory processing," *PloS one,* vol. 7, p. e44560, 2012.
[75] G. Ohtsuki, C. Piochon, and C. Hansel, "Climbing fiber signaling and cerebellar gain control," *Frontiers in cellular neuroscience,* vol. 3, p. 4, 2009.
[76] R. Brooks, "A robust layered control system for a mobile robot," *IEEE Journal on Robotics and Automation,* vol. 2, pp. 14-23, 1986.
[77] R. A. Brooks, *Cambrian intelligence: the early history of the new AI*. Cambridge, Mass. ; London: MIT Press, 1999.
[78] S. J. Blakemore, D. Wolpert, and C. Frith, "Why can't you tickle yourself?," *NeuroReport,* vol. 11, pp. R11-R16, 2000.
[79] O. Baumann, R. Borra, J. Bower, K. Cullen, C. Habas, R. Ivry*, et al.*, "Consensus Paper: The Role of the Cerebellum in Perceptual Processes," *Cerebellum,* vol. 14, pp. 197-220, 2015.
[80] K. Doya, "What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?," *Neural Networks,* vol. 12, pp. 961-974, 1999.
[81] R. S. Snider and A. Stowell, "Receiving areas of the tactile, auditory and visual systems in the cerebellum," *Journal of Neurophysiology,* vol. 7, pp. 331-357, 1944.
[82] D. Oertel and E. D. Young, "What's a cerebellar circuit doing in the auditory system?," *Trends in Neurosciences,* vol. 27, pp. 104-110, 2004.
[83] A. Petacchi, A. R. Laird, P. T. Fox, and J. M. Bower, "Cerebellum and auditory function: An ALE meta-analysis of functional neuroimaging studies," *Human Brain Mapping,* vol. 25, pp. 118-128, 2005.

[84] P. Maria Sens and C. I. Ribeiro de Almeida, "Participation of the Cerebellum in Auditory Processing," *Brazilian Journal of Otorhinolaryngology,* vol. 73, pp. 266-270, 2007.
[85] S. Singla, "A Cerebellum-like Circuit in the Auditory System Cancels Self-Generated Sounds," 2016.
[86] N. M. McLachlan and S. J. Wilson, "The Contribution of Brainstem and Cerebellar Pathways to Auditory Recognition," *Frontiers in psychology,* vol. 8, p. 265, 2017.
[87] K. Mathiak, I. Hertrich, W. Grodd, and H. Ackermann, "Discrimination of temporal information at the cerebellum: functional magnetic resonance imaging of nonverbal auditory memory," *Neuroimage,* vol. 21, pp. 154-162, 2004.
[88] M. Schwartze, A. Tavano, E. Schröger, and S. A. Kotz, "Temporal aspects of prediction in audition: Cortical and subcortical neural mechanisms," *International Journal of Psychophysiology,* vol. 83, pp. 200-207, 2012.
[89] R. B. Ivry, R. M. Spencer, H. N. Zelaznik, and J. Diedrichsen, "The Cerebellum and Event Timing," *Annals of the New York Academy of Sciences,* vol. 978, pp. 302-317, 2002.
[90] S. Wilkinson and B. Alderson-Day, "Voices and Thoughts in Psychosis: An Introduction," *Review of Philosophy and Psychology,* vol. 7, pp. 529-540, 2016.
[91] S. R. Anderson, M. J. Pearson, A. Pipe, T. Prescott, P. Dean, and J. Porrill, "Adaptive Cancelation of Self-Generated Sensory Signals in a Whisking Robot," *Robotics, IEEE Transactions on,* vol. 26, pp. 1065-1076, 2010.
[92] E. D'Angelo, A. Antonietti, S. Casali, C. Casellato, J. A. Garrido, N. R. Luque*, et al.*, "Modeling the Cerebellar Microcircuit: New Strategies for a Long-Standing Issue," *Frontiers in cellular neuroscience,* vol. 10, p. 176, 2016.
[93] R. D. Pinzon Morales and Y. Hirata, "Evaluation of Teaching Signals for Motor Control in the Cerebellum during Real-World Robot Application," *Brain sciences,* vol. 6, p. 62, 2016.
[94] S. S. D. Widrow B, *Adaptive Signal Processing*. ENGLEWOOD CLIFFS U6 - ctx_ver=Z39.88-2004&ctx_enc=info%3Aofi%2Fenc%3AUTF-8&rfr_id=info%3Asid%2Fsummon.serialssolutions.com&rft_val_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx%3Abook&rft.genre=book&rft.title=ADAPTIVE+SIGNAL+PROCESSING&rft.au=WIDROW+B+%2C+STEARNS+S+D&rft.date=1985&rft.pub=PRENTICE+HALL&rft.externalDBID=EPFWF&rft.externalDocID=B0165660&paramdict=en-US U7 - Book: PRENTICE HALL, 1985.
[95] D. Marr, "A theory of cerebellar cortex," *The Journal of Physiology,* vol. 202, pp. 437-470.1, 1969.
[96] J. S. Albus, "A theory of cerebellar function," *Mathematical Biosciences,* vol. 10, pp. 25-61, 1971.
[97] P. Dean and J. Porrill, "The cerebellum as an adaptive filter: a general model?," *Functional neurology,* vol. 25, p. 173, 2010.
[98] P. Dean and J. Porrill, "Evaluating the adaptive-filter model of the cerebellum," *The Journal of Physiology,* vol. 589, pp. 3459-3470, 2011.
[99] P. Dean and J. Porrill, "Adaptive-filter Models of the Cerebellum: Computational Analysis," *The Cerebellum,* vol. 7, pp. 567-571, 2008.
[100] J. Porrill and P. Dean, "Silent synapses, LTP, and the indirect parallel-fibre pathway: computational consequences of optimal cerebellar noise-processing," *PLoS computational biology,* vol. 4, p. e1000085, 2008.

[101] P. Bratby, J. Sneyd, and J. Montgomery, "Computational Architecture of the Granular Layer of Cerebellum-Like Structures," *The Cerebellum,* vol. 16, pp. 15-25, 2017.
[102] C. Hofst, #246, tter, M. Gil, K. Eng, G. Indiveri*, et al.*, "The cerebellum chip: an analog VLSI implementation of a cerebellar model of classical conditioning," presented at the Proceedings of the 17th International Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, 2004.
[103] M. Mikaitis, G. Pineda García, J. C. Knight, and S. B. Furber, "Neuromodulated Synaptic Plasticity on the SpiNNaker Neuromorphic System," *Frontiers in neuroscience,* vol. 12, p. 105, 2018.
[104] M. Kawato, "Internal models for motor control and trajectory planning," *Current Opinion in Neurobiology,* vol. 9, pp. 718-727, 12/1/ 1999.
[105] M. Kawato, K. Furukawa, and R. Suzuki, "A hierarchical neural-network model for control and learning of voluntary movement," *Biol Cybern,* vol. 57, pp. 169-85, 1987.
[106] M. I. Jordan and D. E. Rumelhart, "Forward models: Supervised learning with a distal teacher," *Cognitive Science,* vol. 16, pp. 307-354, 1992.
[107] M. Kawato, K. Furukawa, and R. Suzuki, "A hierarchical neural-network model for control and learning of voluntary movement," *Biological cybernetics,* vol. 57, pp. 169-185, 1987.
[108] H. Gomi and M. Kawato, "Neural network control for a closed-loop System using Feedback-error-learning," *Neural Networks,* vol. 6, pp. 933-946, 1993.
[109] B. E. Stein and T. R. Stanford, "Multisensory integration: current issues from the perspective of the single neuron," *Nature Reviews Neuroscience,* vol. 9, pp. 255-266, 2008.
[110] F. R. Robinson and A. F. Fuchs, "The Role of the Cerebellum in Voluntary Eye Movements," *Annual review of neuroscience,* vol. 24, pp. 981-1004, 2001.
[111] T. J. Sejnowski, "Storing covariance with nonlinearly interacting neurons," *J Math Biol,* vol. 4, pp. 303-21, Oct 20 1977.
[112] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems,* vol. 96, pp. 184-210, 2017.
[113] M. D. Baxendale, M. J. Pearson, M. Nibouche, E. L. Secco, and A. G. Pipe, "Self-adaptive Context Aware Audio Localization for Robots Using Parallel Cerebellar Models," in *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Guildford, UK, July 19–21, 2017, Proceedings*, Y. Gao, S. Fallah, Y. Jin, and C. Lekakou, Eds., ed Cham: Springer International Publishing, 2017, pp. 66-78.
[114] L. H. Crockett, R. Elliot, M. Enderwitz, and R. Stewart, *The Zynq Book*, 2014.
[115] R. Yan, T. Rodemann, and B. Wrede, "Computational Audiovisual Scene Analysis in Online Adaptation of Audio-Motor Maps," *IEEE Transactions on Autonomous Mental Development,* vol. 5, pp. 273-287, 2013.
[116] Student, "The Probable Error of a Mean," *Biometrika,* vol. 6, pp. 1-25, 1908.
[117] D. M. Wolpert, Z. Ghahramani, and M. I. Jordan, "An Internal Model for Sensorimotor Integration," *Science,* vol. 269, pp. 1880-1882, 1995.
[118] R. C. Miall and D. M. Wolpert, "Forward Models for Physiological Motor Control," *Neural Networks,* vol. 9, pp. 1265-1279, 1996.
[119] M. Kawato and D. Wolpert, "Internal models for motor control," *Novartis Found Symp,* vol. 218, pp. 291-304; discussion 304-7, 1998.
[120] D. M. Wolpert, R. C. Miall, and M. Kawato, "Internal models in the cerebellum," *Trends in Cognitive Sciences,* vol. 2, pp. 338-347, 1998.

[121] H. Imamizu and M. Kawato, "Cerebellar Internal Models: Implications for the Dexterous Use of Tools," *Cerebellum (London, England),* vol. 11, pp. 325-335, 2012.
[122] M. Haruno, D. M. Wolpert, and M. Kawato, "MOSAIC Model for Sensorimotor Learning and Control," *Neural Computation,* vol. 13, pp. 2201-2220, 2001.
[123] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.,* vol. 3, pp. 79-87, 1991.
[124] K. S. Narendra, J. Balakrishnan, and M. K. Ciliz, "Adaptation and learning using multiple models, switching, and tuning," vol. 15, ed: IEEE, 1995, pp. 37-51.
[125] Z. Ghahramani and D. M. Wolpert, "Modular decomposition in visuomotor learning," *Nature,* vol. 386, pp. 392-395, 1997.
[126] K. S. Narendra and J. Balakrishnan, "Adaptive control using multiple models," *IEEE Transactions on Automatic Control,* vol. 42, pp. 171-187, 1997.
[127] P. Vetter and D. M. Wolpert, "Context estimation for sensorimotor control," *Journal of neurophysiology,* vol. 84, p. 1026, 2000.
[128] K. S. Narendra and Z. Han, "The changing face of adaptive control: The use of multiple models," *Annual Reviews in Control,* vol. 35, p. 1, 2011.
[129] K. S. Narendra and Z. Han, "A new approach to adaptive control using multiple models," *International Journal of Adaptive Control and Signal Processing,* vol. 26, pp. 778-799, 2012.
[130] E. Wilson, S. Anderson, T. Assaf, M. Pearson, P. Walters, T. Prescott*, et al.*, "Bioinspired Control of Electro-Active Polymers for Next Generation Soft Robots," *Advances in Autonomous Robotics,* vol. 7429, pp. 424-425, 2012.
[131] Y. Demiris and B. Khadhouri, "Hierarchical attentive multiple models for execution and recognition of actions," *Robotics and Autonomous Systems,* vol. 54, pp. 361-369, 2006.
[132] M. Haruno, D. M. Wolpert, and M. Kawato, "Hierarchical MOSAIC for movement generation," *International Congress Series,* vol. 1250, pp. 575-590, 2003.
[133] N. Sugimoto, M. Haruno, K. Doya, and M. Kawato, "MOSAIC for Multiple-Reward Environments," *Neural Computation,* vol. 24, pp. 577-606, 2012.
[134] M. Kawato and H. Gomi, "The cerebellum and VOR/OKR learning models," *Trends Neurosci,* vol. 15, pp. 445-53, Nov 1992.
[135] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE,* vol. 77, pp. 257-286, 1989.
[136] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic Scene Classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine,* vol. 32, pp. 16-34, 2015.
[137] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and Classification of Acoustic Scenes and Events," *IEEE Transactions on Multimedia,* vol. 17, pp. 1733-1746, 2015.
[138] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," 2016, pp. 6445-6449.
[139] T. Giannakopoulos and A. Pikrakis, *Introduction to audio analysis: a MATLAB approach*, First ed. Amsterdam: Academic Press, 2014.
[140] K. Mehmed, "Data Reduction," in *Data Mining:Concepts, Models, Methods, and Algorithms*, ed: Wiley-IEEE Press, 2003, p. 360.
[141] A. R. Webb, *Statistical pattern recognition*, 2nd ed. New York: Wiley, 2002.

[142] A. Tidemann and P. Ozturk, "A self-organising multiple model architecture for motor imitation," *International journal of intelligent information and database systems,* vol. 4, 2010.
[143] E. Escobar-Juárez, G. Schillaci, J. Hermosillo-Valadez, and B. Lara-Guzmán, "A Self-Organized Internal Models Architecture for Coding Sensory–Motor Schemes," *Frontiers in Robotics and AI,* vol. 3, 2016.
[144] A. Tidemann and P. Öztürk, "Self-organizing Multiple Models for Imitation: Teaching a Robot to Dance the YMCA." vol. 4570, ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 291-302.
[145] Z. Cha, D. Florencio, and Z. Zhengyou, "Why does PHAT work well in lownoise, reverberative environments?," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 2565-2568.
[146] E. Georganti, T. May, S. van de Par, and J. Mourjopoulos, "Extracting Sound-Source-Distance Information from Binaural Signals," in *The Technology of Binaural Listening*, J. Blauert, Ed., ed Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 171-199.

Appendix 1 Selected audio maps and data sets

A1.1 Calibrated versus uncalibrated SSL

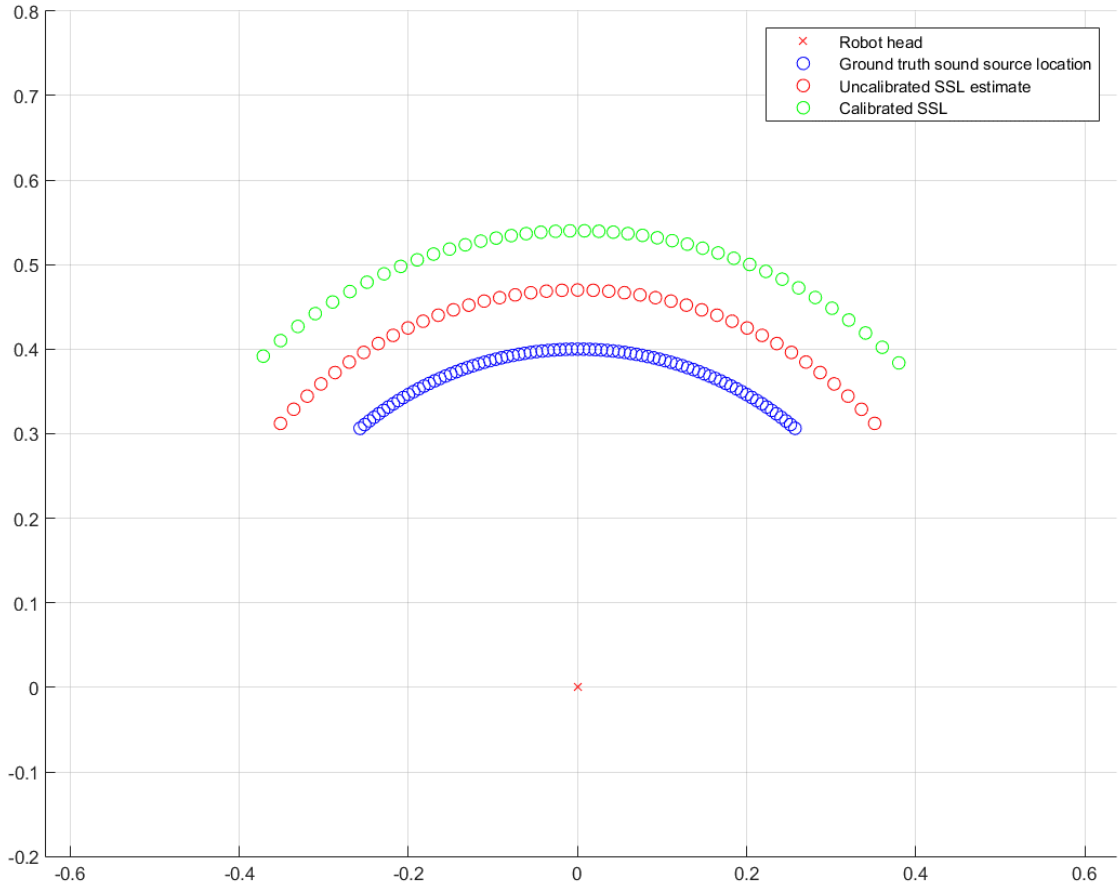


Figure 67. Results of cerebellar calibration post-learning: cerebellar calibration (green circles) versus uncalibrated SSL estimate (red circles). Blue circles are ground truth sound source position. Axes show x,y distance in metres with robot head at the origin.

Table 6. Data set: cerebellar calibration versus uncalibrated SSL.

Ground truth azimuth, degrees	Uncalibrated SLL, degrees	Calibrated SLL, degrees
-40	-48.37	-43.49
-39	-48.37	-43.49
-38	-45.58	-40.58
-37	-45.58	-40.58
-36	-42.86	-37.77
-35	-42.86	-37.77
-34	-42.86	-37.77
-33	-40.21	-35.06
-32	-40.21	-35.06
-31	-37.61	-32.44
-30	-37.61	-32.44
-29	-35.06	-29.90
-28	-35.06	-29.90
-27	-32.55	-27.45
-26	-32.55	-27.45
-25	-30.09	-25.08
-24	-30.09	-25.08
-23	-27.65	-22.79
-22	-27.65	-22.79
-21	-25.25	-20.56
-20	-25.25	-20.56
-19	-22.88	-18.39
-18	-22.88	-18.39
-17	-20.53	-16.28
-16	-20.53	-16.28
-15	-18.20	-14.23
-14	-18.20	-14.23
-13	-15.88	-12.23
-12	-15.88	-12.23
-11	-13.59	-10.27
-10	-11.30	-8.36
-9	-11.30	-8.36
-8	-9.03	-6.47
-7	-9.03	-6.47
-6	-6.77	-4.61
-5	-6.77	-4.61
-4	-4.51	-2.78
-3	-4.51	-2.78
-2	-2.25	-0.96
-1	-2.25	-0.96

Ground truth azimuth, degrees	Uncalibrated SLL, degrees	Calibrated SLL, degrees
0	0.00	0.86
1	0.00	0.86
2	2.25	2.67
3	2.25	2.67
4	4.51	4.49
5	4.51	4.49
6	6.77	6.32
7	6.77	6.32
8	9.03	8.17
9	9.03	8.17
10	11.3	10.05
11	11.3	10.05
12	13.59	11.96
13	15.88	13.90
14	15.88	13.90
15	18.20	15.89
16	18.20	15.89
17	20.53	17.92
18	20.53	17.92
19	22.88	20.00
20	22.88	20.00
21	25.25	22.15
22	25.25	22.15
23	27.65	24.35
24	27.65	24.35
25	30.09	26.62
26	30.09	26.62
27	32.55	28.95
28	32.55	28.95
29	35.06	31.37
30	35.06	31.37
31	37.61	33.86
32	37.61	33.86
33	40.21	36.43
34	40.21	36.43
35	42.86	39.09
36	42.86	39.09
37	45.58	41.85
38	45.58	41.85
39	45.58	41.85
40	48.37	44.71

A1.2 Multiple models calibration versus general single model calibration

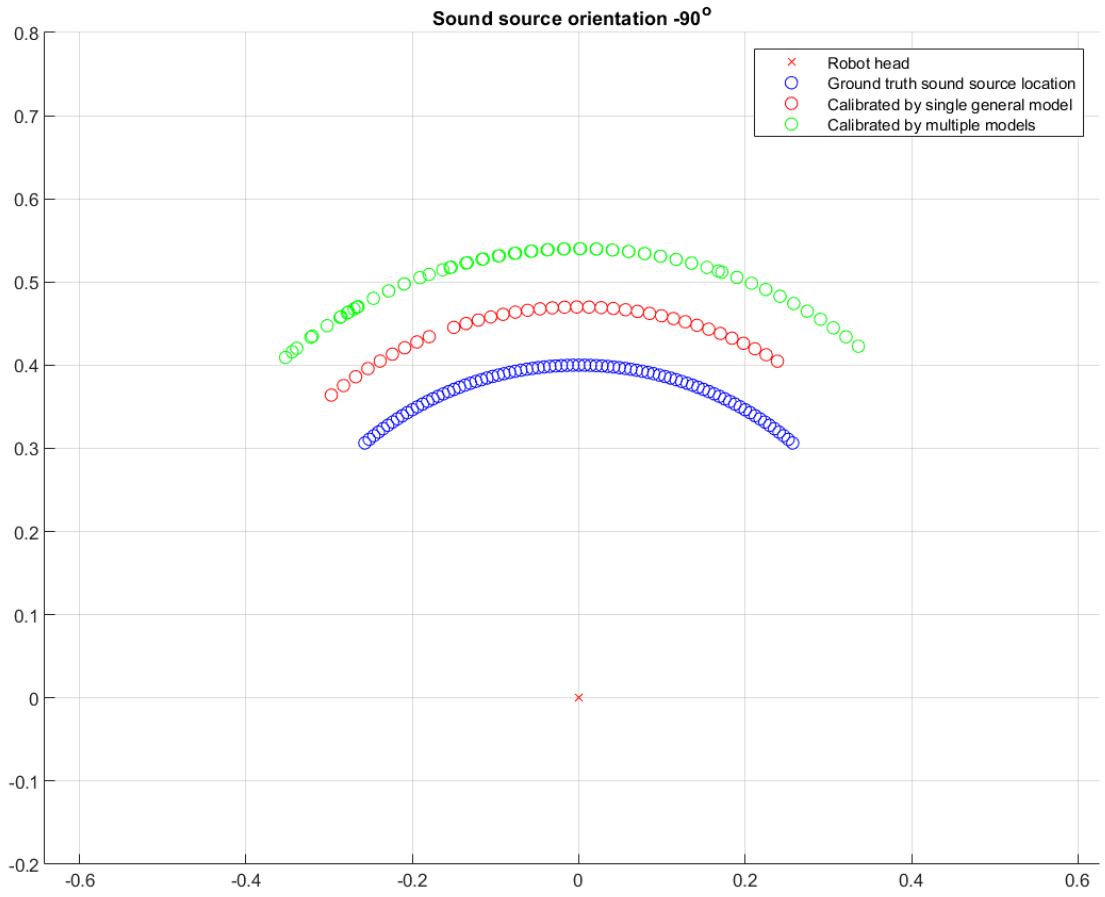


Figure 68. Results of cerebellar calibration post-learning: multiple models (green circles) versus a single general model which has learned in all contexts (red circles). Blue circles are ground truth sound source position. Axes show x,y distance in metres with robot head at the origin. Context is ϕ=-90º.

Table 7. Data set: multiple models versus single general model, ϕ=-90°.

Ground truth azimuth, degrees	Calibrated SLL, single general model, degrees	Calibrated SLL, multiple models, degrees
-40	-36.98	-39.62
-39	-36.98	-38.89
-38	-39.24	-40.72
-37	-39.24	-38.9
-36	-36.98	-36.36
-35	-36.98	-36.59
-34	-34.77	-34.05
-33	-32.62	-32.08
-32	-30.52	-30.83
-31	-30.52	-31.94
-30	-30.52	-31.01
-29	-30.52	-30.01
-28	-30.52	-29.57
-27	-30.52	-29.44
-26	-30.52	-29.42
-25	-28.46	-27.19
-24	-26.43	-25.01
-23	-26.43	-25.04
-22	-24.44	-22.84
-21	-24.44	-22.87
-20	-22.48	-20.69
-19	-18.64	-16.47
-18	-18.64	-19.46
-17	-18.64	-17.61
-16	-18.64	-16.65
-15	-16.74	-14.36
-14	-16.74	-14.53
-13	-14.87	-12.26
-12	-14.87	-12.43
-11	-13.01	-10.16
-10	-13.01	-10.32
-9	-11.16	-8.07
-8	-11.16	-8.22
-7	-9.33	-5.99
-6	-9.33	-6.13
-5	-7.51	-3.91
-4	-7.51	-4.04
-3	-5.69	-1.84
-2	-5.69	-1.95
-1	-3.89	0.23

Ground truth azimuth, degrees	Calibrated SLL, single general model, degrees	Calibrated SLL, multiple models, degrees
0	-3.89	0.13
1	-2.09	2.29
2	-2.09	2.20
3	-0.29	4.35
4	-0.29	4.28
5	1.50	6.40
6	1.50	6.34
7	3.29	8.45
8	3.29	8.40
9	5.07	10.49
10	5.07	10.45
11	6.86	12.53
12	6.86	12.49
13	8.64	14.56
14	8.64	14.53
15	10.43	16.58
16	12.22	18.58
17	12.22	18.10
18	12.22	18.58
19	14.01	20.61
20	14.01	20.60
21	14.01	20.61
22	15.80	22.61
23	17.61	24.61
24	17.61	24.60
25	19.42	26.60
26	19.42	26.60
27	21.23	28.59
28	21.23	28.58
29	23.06	30.57
30	23.06	30.57
31	24.91	32.55
32	24.91	32.54
33	24.91	32.55
34	26.76	34.52
35	26.76	34.52
36	26.76	34.52
37	28.64	36.50
38	28.64	36.50
39	28.64	36.50
40	30.53	38.48

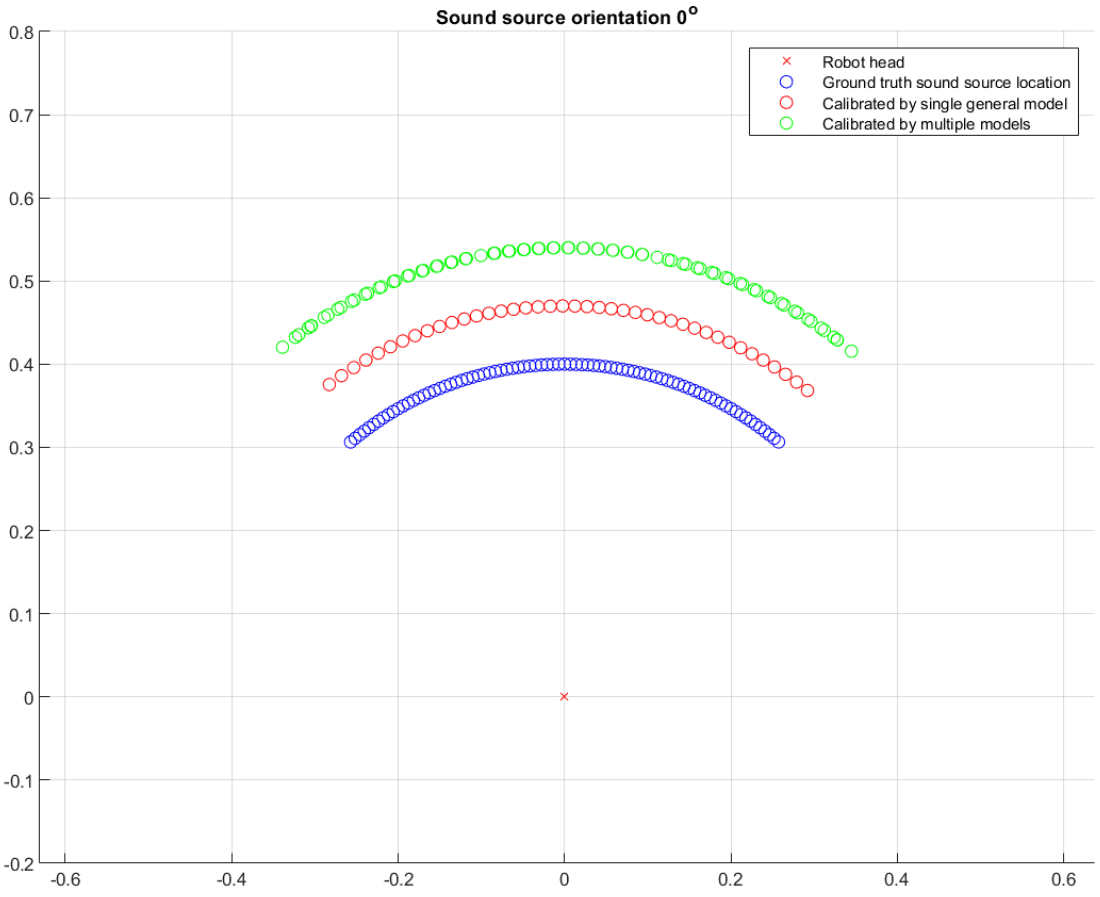


Figure 69. Results of cerebellar calibration post-learning: multiple models (green circles) versus a single general model which has learned in all contexts (red circles). Blue circles are ground truth sound source position. Axes show x,y distance in metres with robot head at the origin. Contexts is $\phi=0^{o}$.

Table 8. Data set: multiple models versus single general model, $\phi=0^o$.

Ground truth azimuth, degrees	Calibrated SLL, single general model, degrees	Calibrated SLL, multiple models, degrees
-40	-36.98	-36.27
-39	-36.98	-38.89
-38	-34.77	-36.29
-37	-34.77	-36.81
-36	-32.62	-34.28
-35	-32.62	-34.77
-34	-32.62	-34.36
-33	-30.52	-31.79
-32	-30.52	-32.33
-31	-28.46	-29.86
-30	-28.46	-30.31
-29	-26.43	-27.93
-28	-26.43	-28.31
-27	-24.44	-26.00
-26	-24.44	-26.32
-25	-22.48	-24.08
-24	-22.48	-24.35
-23	-20.55	-22.16
-22	-20.55	-22.39
-21	-18.64	-20.24
-20	-18.64	-20.44
-19	-16.74	-18.33
-18	-16.74	-18.50
-17	-14.87	-16.42
-16	-14.87	-16.58
-15	-13.01	-14.52
-14	-13.01	-14.66
-13	-11.16	-12.62
-12	-11.16	-12.74
-11	-9.33	-10.72
-10	-7.51	-8.94
-9	-7.51	-9.08
-8	-5.69	-7.04
-7	-5.69	-7.17
-6	-3.89	-5.14
-5	-3.89	-5.26
-4	-2.09	-3.25
-3	-2.09	-3.36
-2	-0.29	-1.36
-1	-0.29	-1.46

Ground truth azimuth, degrees	Calibrated SLL, single general model, degrees	Calibrated SLL, multiple models, degrees
0	1.50	0.53
1	1.50	0.44
2	3.29	2.42
3	3.29	2.34
4	5.07	4.31
5	5.07	4.23
6	6.86	6.20
7	6.86	6.13
8	8.64	8.10
9	8.64	8.02
10	10.43	9.99
11	10.43	9.92
12	12.22	11.89
13	14.01	13.73
14	14.01	13.35
15	15.80	15.64
16	15.80	15.28
17	17.61	17.55
18	17.61	17.21
19	19.42	19.47
20	19.42	19.14
21	21.23	21.40
22	21.23	21.07
23	23.06	23.34
24	23.06	23.01
25	24.91	25.29
26	24.91	24.95
27	26.76	27.26
28	26.76	26.91
29	28.64	29.24
30	28.64	28.87
31	30.53	31.25
32	30.53	30.84
33	32.45	33.27
34	32.45	32.83
35	34.40	35.31
36	34.40	34.82
37	36.38	37.39
38	36.38	36.84
39	36.38	37.32
40	38.40	39.67

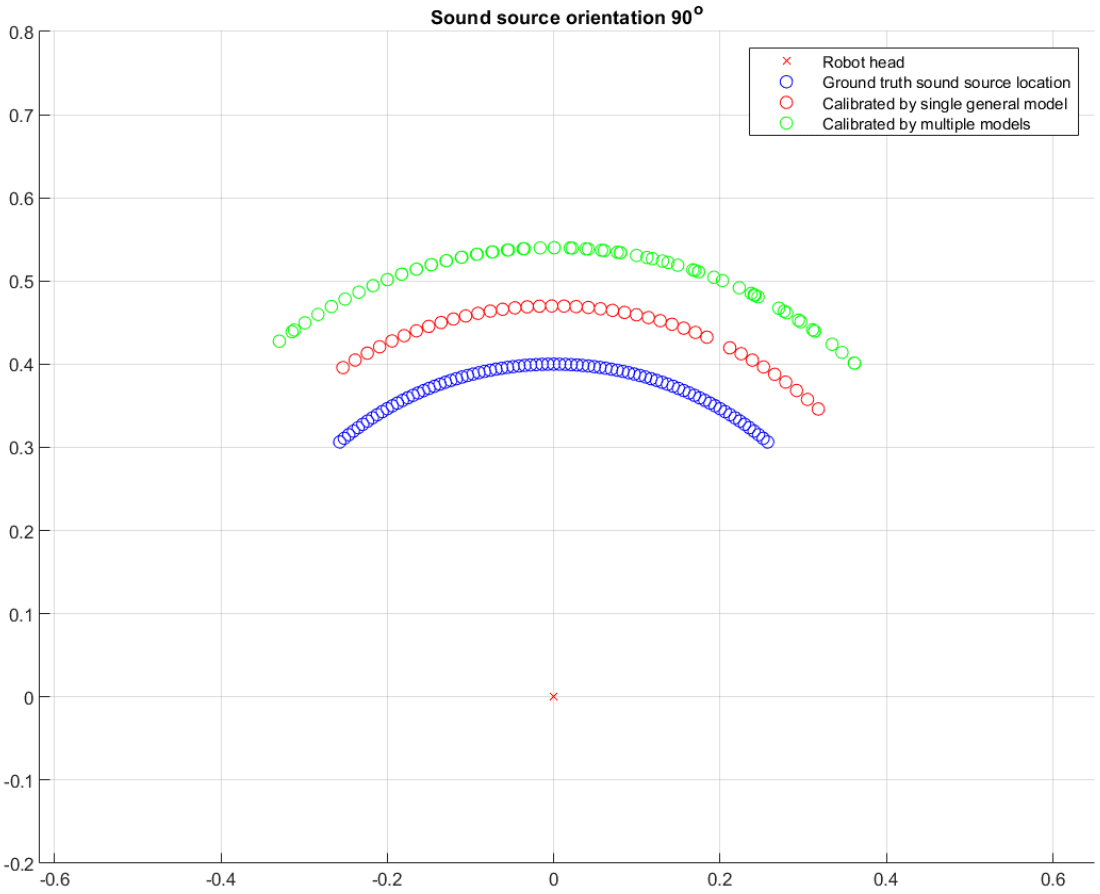


Figure 70. Results of cerebellar calibration post-learning: multiple models (green circles) versus a single general model which has learned in all contexts (red circles). Blue circles are ground truth sound source position. Axes show x,y distance in metres with robot head at the origin. Contexts is ϕ=90º.

Table 9. Data set: multiple models versus single general model, $\phi=90^o$.

Ground truth azimuth, degrees	Calibrated SLL, single general model, degrees	Calibrated SLL, multiple models, degrees
-40	-32.62	-35.25
-39	-30.52	-37.63
-38	-30.52	-37.63
-37	-28.46	-35.62
-36	-28.46	-35.62
-35	-26.43	-33.62
-34	-26.43	-33.62
-33	-24.44	-31.63
-32	-24.44	-31.63
-31	-22.48	-29.65
-30	-22.48	-29.65
-29	-20.55	-27.67
-28	-20.55	-27.68
-27	-18.64	-25.7
-26	-18.64	-25.7
-25	-18.64	-25.7
-24	-16.74	-23.7
-23	-16.74	-23.73
-22	-14.87	-21.72
-21	-14.87	-21.76
-20	-13.01	-19.74
-19	-13.01	-19.78
-18	-11.16	-17.76
-17	-11.16	-17.81
-16	-9.33	-15.78
-15	-9.33	-15.84
-14	-7.51	-13.79
-13	-7.51	-13.86
-12	-5.69	-11.79
-11	-5.69	-11.88
-10	-3.89	-9.79
-9	-3.89	-9.89
-8	-2.09	-7.79
-7	-2.09	-7.9
-6	-0.29	-5.77
-5	-0.29	-5.91
-4	1.5	-3.75
-3	1.5	-3.92
-2	3.29	-1.72
-1	5.07	0.08
0	5.07	0.05
1	5.07	0.08
2	6.86	2.35
3	6.86	2.09
4	8.64	4.40
5	8.64	4.10
6	10.43	6.45
7	10.43	6.12
8	12.22	8.50
9	12.22	8.14
10	12.22	8.56
11	14.01	11.98
12	14.01	10.62
13	15.80	14.01
14	15.80	12.68
15	17.61	16.03
16	17.61	14.73
17	19.42	18.04
18	21.23	18.80
19	21.23	18.32
20	21.23	18.84
21	23.06	22.05
22	23.06	20.89
23	26.76	26.07
24	26.76	24.39
25	28.64	26.61
26	28.64	26.48
27	28.64	26.61
28	28.64	27.06
29	30.53	30.07
30	32.45	31.25
31	32.45	30.85
32	34.40	33.37
33	34.40	33.00
34	36.38	35.51
35	36.38	35.19
36	36.38	35.46
37	38.40	38.28
38	40.46	39.92
39	42.58	42.04
40	42.58	41.98

A1.3 Multiple models with and without RP

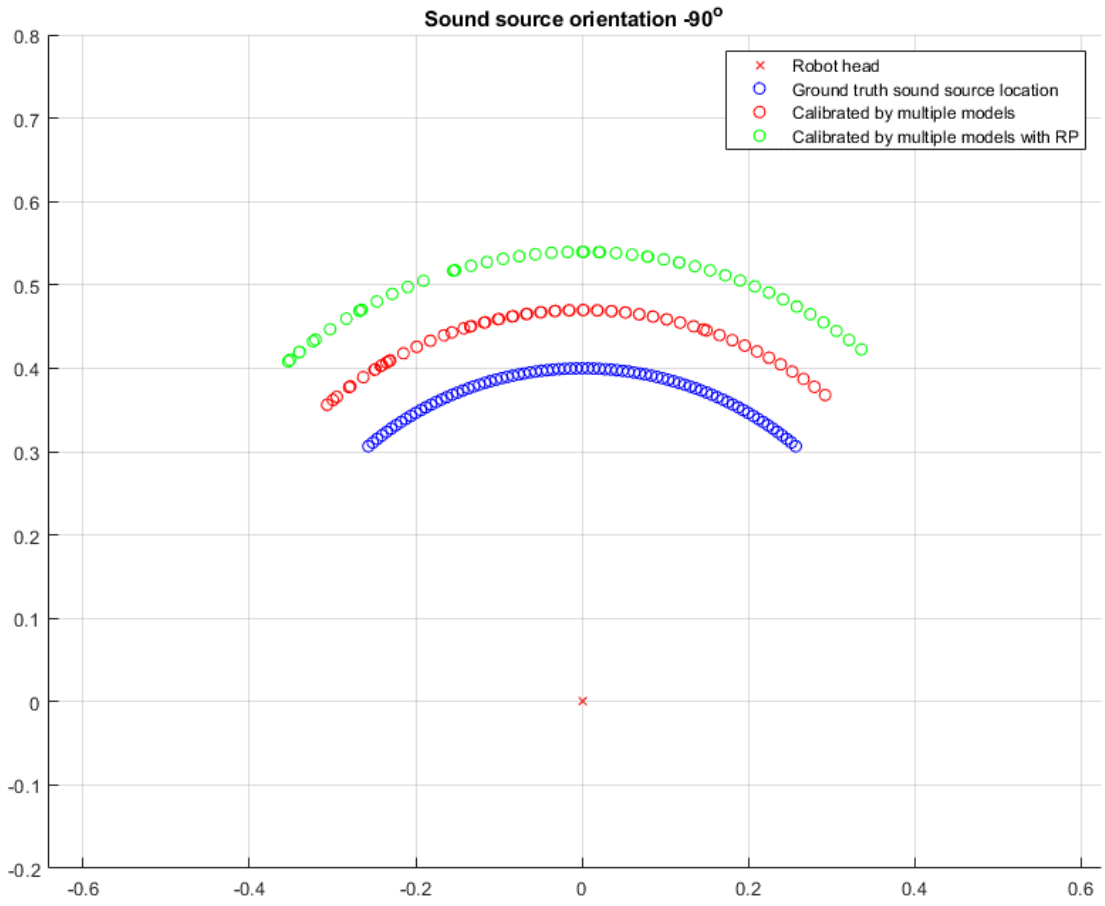


Figure 71. Results of cerebellar calibration post-learning: multiple models with RP (green circles) versus multiple models without RP (red circles). Blue circles are ground truth sound source position. Axes show x,y distance in metres with robot head at the origin. Context is ϕ=-90º.

Table 10. Data set: multiple models versus multiple models with RP, ϕ=-90^o.

Ground truth azimuth, degrees	Calibrated SLL, multiple models, degrees	Calibrated SLL, multiple models with RP, degrees
-40	-39.62	-40.58
-39	-38.89	-38.98
-38	-40.72	-40.86
-37	-38.9	-39.06
-36	-36.36	-36.43
-35	-36.59	-36.81
-34	-34.05	-34.14
-33	-32.08	-31.67
-32	-30.83	-29.47
-31	-31.94	-29.74
-30	-31.01	-29.49
-29	-30.01	-29.43
-28	-29.57	-29.41
-27	-29.44	-29.41
-26	-29.42	-29.41
-25	-27.19	-27.19
-24	-25.01	-25.00
-23	-25.04	-25.00
-22	-22.84	-22.84
-21	-22.87	-22.84
-20	-20.69	-20.69
-19	-16.47	-16.44
-18	-19.46	-16.68
-17	-17.61	-16.47
-16	-16.65	-16.44
-15	-14.36	-14.33
-14	-14.53	-14.33
-13	-12.26	-12.23
-12	-12.43	-12.23
-11	-10.16	-10.14
-10	-10.32	-10.14
-9	-8.07	-8.05
-8	-8.22	-8.06
-7	-5.99	-5.97
-6	-6.13	-5.98
-5	-3.91	-3.9
-4	-4.04	-3.9
-3	-1.84	-1.85
-2	-1.95	-1.83
-1	0.23	0.22

Ground truth azimuth, degrees	Calibrated SLL, multiple models, degrees	Calibrated SLL, multiple models with RP, degrees
0	0.13	0.04
1	2.29	2.28
2	2.20	2.13
3	4.35	4.34
4	4.28	4.35
5	6.40	6.41
6	6.34	6.40
7	8.45	8.45
8	8.40	8.35
9	10.49	10.49
10	10.45	10.49
11	12.53	12.53
12	12.49	12.46
13	14.56	14.56
14	14.53	14.56
15	16.58	16.58
16	18.58	18.6
17	18.1	18.59
18	18.58	18.6
19	20.61	20.61
20	20.6	20.61
21	20.61	20.61
22	22.61	22.61
23	24.61	24.61
24	24.6	24.59
25	26.6	26.6
26	26.6	26.59
27	28.59	28.59
28	28.58	28.58
29	30.57	30.57
30	30.57	30.56
31	32.55	32.55
32	32.54	32.54
33	32.55	32.55
34	34.52	34.52
35	34.52	34.52
36	34.52	34.52
37	36.5	36.5
38	36.5	36.5
39	36.5	36.5
40	38.48	38.48

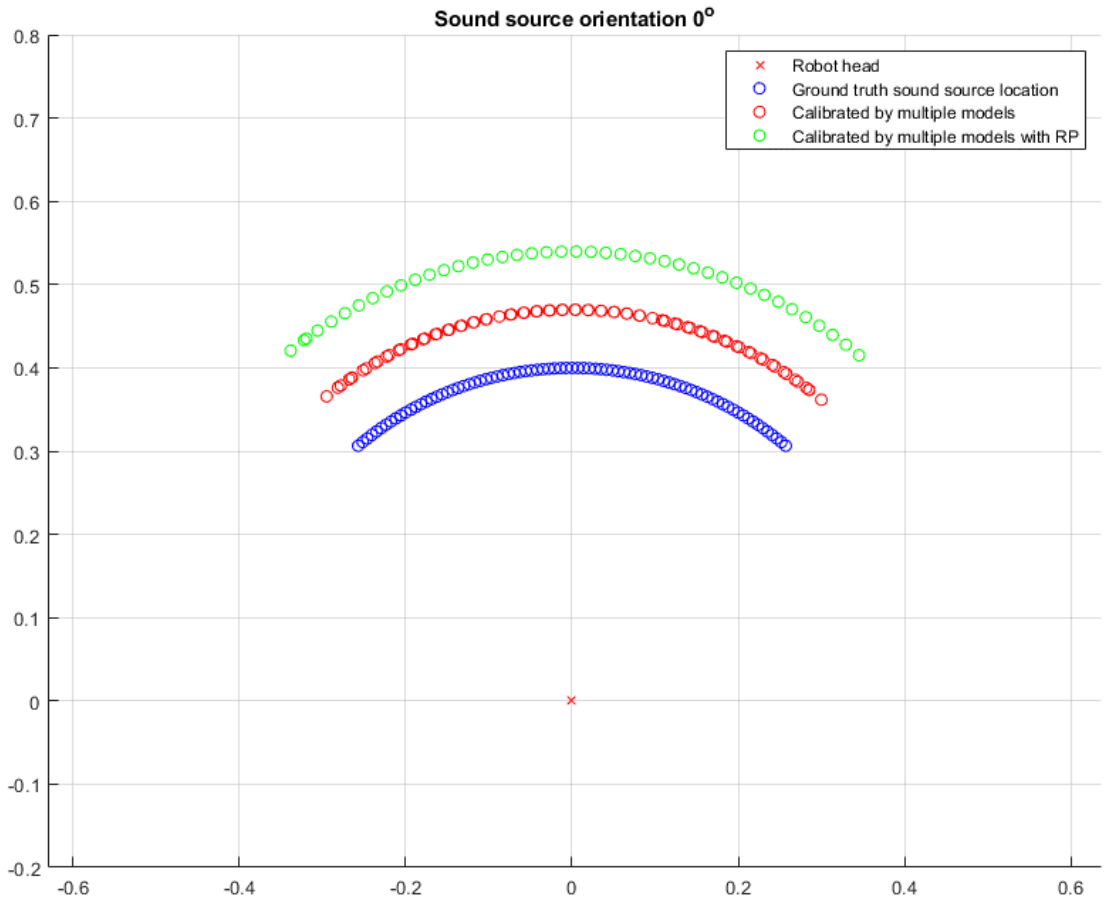


Figure 72. Results of cerebellar calibration post-learning: multiple models with RP (green circles) versus multiple models without RP (red circles). Blue circles are ground truth sound source position. Axes show x,y distance in metres with robot head at the origin. Context is ϕ=0º.

Table 11. Data set: multiple models versus multiple models with RP, $\phi=0^o$.

Ground truth azimuth, degrees	Calibrated SLL, multiple models, degrees	Calibrated SLL, multiple models with RP, degrees
-40	-36.27	-36.27
-39	-38.89	-38.81
-38	-36.29	-36.64
-37	-36.81	-36.64
-36	-34.28	-34.52
-35	-34.77	-34.52
-34	-34.36	-34.52
-33	-31.79	-32.43
-32	-32.33	-32.43
-31	-29.86	-30.37
-30	-30.31	-30.37
-29	-27.93	-28.35
-28	-28.31	-28.35
-27	-26.00	-26.34
-26	-26.32	-26.34
-25	-24.08	-24.36
-24	-24.35	-24.36
-23	-22.16	-22.39
-22	-22.39	-22.39
-21	-20.24	-20.44
-20	-20.44	-20.44
-19	-18.33	-18.5
-18	-18.5	-18.5
-17	-16.42	-16.57
-16	-16.58	-16.57
-15	-14.52	-14.65
-14	-14.66	-14.65
-13	-12.62	-12.74
-12	-12.74	-12.74
-11	-10.72	-10.83
-10	-8.94	-8.93
-9	-9.08	-8.93
-8	-7.04	-7.03
-7	-7.17	-7.03
-6	-5.14	-5.14
-5	-5.26	-5.14
-4	-3.25	-3.24
-3	-3.36	-3.25
-2	-1.36	-1.35
-1	-1.46	-1.35

Ground truth azimuth, degrees	Calibrated SLL, multiple models, degrees	Calibrated SLL, multiple models with RP, degrees
0	0.53	0.54
1	0.44	0.54
2	2.42	2.43
3	2.34	2.42
4	4.31	4.32
5	4.23	4.32
6	6.20	6.21
7	6.13	6.21
8	8.10	8.10
9	8.02	8.10
10	9.99	10.00
11	9.92	10.00
12	11.89	11.90
13	13.73	13.80
14	13.35	13.80
15	15.64	15.71
16	15.28	15.71
17	17.55	17.63
18	17.21	17.62
19	19.47	19.55
20	19.14	19.55
21	21.4	21.48
22	21.07	21.48
23	23.34	23.43
24	23.01	23.42
25	25.29	25.39
26	24.95	25.37
27	27.26	27.36
28	26.91	27.36
29	29.24	29.35
30	28.87	29.35
31	31.25	31.37
32	30.84	31.37
33	33.27	33.41
34	32.83	33.41
35	35.31	35.49
36	34.82	35.49
37	37.39	37.60
38	36.84	37.60
39	37.32	37.60
40	39.67	39.75

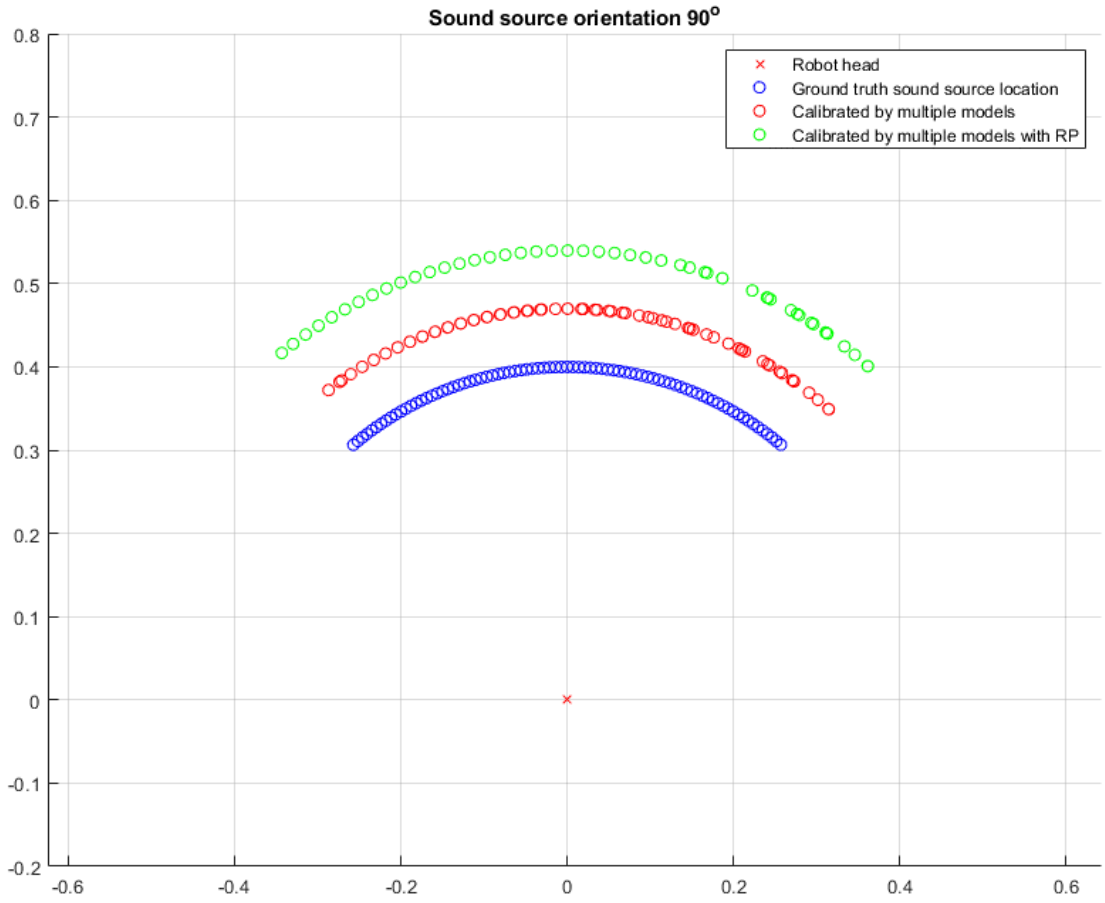


Figure 73. Results of cerebellar calibration post-learning: multiple models with RP (green circles) versus multiple models without RP (red circles). Blue circles are ground truth sound source position. Axes show x,y distance in metres with robot head at the origin. Context is ϕ=90°.

Table 12. Data set: multiple models versus multiple models with RP, ϕ=90°.

Ground truth azimuth, degrees	Calibrated SLL, multiple models, degrees	Calibrated SLL, multiple models with RP, degrees
-40	-35.25	-39.47
-39	-37.63	-37.63
-38	-37.63	-37.63
-37	-35.62	-35.62
-36	-35.62	-35.62
-35	-33.62	-33.62
-34	-33.62	-33.62
-33	-31.63	-31.63
-32	-31.63	-31.63
-31	-29.65	-29.65
-30	-29.65	-29.65
-29	-27.67	-27.68
-28	-27.68	-27.68
-27	-25.70	-25.70
-26	-25.70	-25.70
-25	-25.70	-25.70
-24	-23.70	-23.73
-23	-23.73	-23.73
-22	-21.72	-21.76
-21	-21.76	-21.76
-20	-19.74	-19.79
-19	-19.78	-19.79
-18	-17.76	-17.82
-17	-17.81	-17.82
-16	-15.78	-15.84
-15	-15.84	-15.84
-14	-13.79	-13.86
-13	-13.86	-13.87
-12	-11.79	-11.89
-11	-11.88	-11.89
-10	-9.79	-9.90
-9	-9.89	-9.90
-8	-7.79	-7.92
-7	-7.90	-7.92
-6	-5.77	-5.93
-5	-5.91	-5.93
-4	-3.75	-3.94
-3	-3.92	-3.94
-2	-1.72	-1.95
-1	0.08	0.04

Ground truth azimuth, degrees	Calibrated SLL, multiple models, degrees	Calibrated SLL, multiple models with RP, degrees
0	0.05	0.04
1	0.08	0.04
2	2.35	2.04
3	2.09	2.04
4	4.40	4.04
5	4.10	4.04
6	6.45	6.05
7	6.12	6.05
8	8.50	8.06
9	8.14	8.05
10	8.56	8.06
11	11.98	10.09
12	10.62	10.07
13	14.01	12.10
14	12.68	12.08
15	16.03	15.83
16	14.73	14.62
17	18.04	17.86
18	18.80	18.17
19	18.32	18.17
20	18.84	18.17
21	22.05	20.24
22	20.89	20.22
23	26.07	24.37
24	24.39	24.35
25	26.61	26.44
26	26.48	26.47
27	26.61	26.58
28	27.06	26.95
29	30.07	29.90
30	31.25	31.15
31	30.85	30.82
32	33.37	33.27
33	33.00	32.97
34	35.51	35.43
35	35.19	35.16
36	35.46	35.38
37	38.28	38.14
38	39.92	39.86
39	42.04	42.02
40	41.98	42.07