UNIVERSITY OF GENOVA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

ITALIAN INSTITUTE OF TECHNOLOGY

ROBOTICS, BRAIN, AND COGNITIVE SCIENCE DEPARTMENT

Towards Multimodal Cognitive Architecture for Human-Robot Shared Perception

by

Omar Khaled Elsayed Mohamed Eldardeer

Thesis submitted for the degree of *Doctor of Philosophy* (35° cycle)
May 2023

Dr. Francesco Rea	Supervisor
Prof. Giulio Sandini	Supervisor
Prof. Paolo Massobrio	Head of the PhD program

Thesis Jury:

Prof. Angelo Cangelosi, *University of Manchester*	External Reviewer/ Examiner
Prof. Angelica Lim, *Simon Fraser University*	External Reviewer
Prof. Dimitri Ognibene, *University of Milano-Bicocca*	External Examiner
Prof. Fulvio Mastrogiovanni, *University of Genova*	Internal Examiner

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

I would like to dedicate this thesis to my mother (the teacher), my mother (the friend), my mother (the inspiration), and my father (the backbone)

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Omar Khaled Elsayed Mohamed Eldardeer
May 2023

Acknowledgements

بسم الله الرحمن الرحيم

قبل كل شيء أشكر الله سبحانه وتعالى على إتمام هذه الرسالة وهذه المرحلة من رحلتي العلمية. وبعدها أتقدم بخالص الشكر والتقدير لأمي وأبي الدكتورة أمل والدكتور خالد فقد أعطوني بسخاء، وعلموني بمعرفة، وأحسنوا في كل التفاصيل منذ ولادتي وحتى الآن. دعمتموني دائما دون تردد بكل ما تملكون. لم ولن أستطع ان أوفي حقوقكم فسامحوني على تقصيري. اللهم فبرحمتك وكرمك فأسعدهم وارض عنهم. كما اود ان اشكر اخواتي رقيه وتقى، الرقيقات والحنونات، صاحبات القلوب الواسعة على تشجيعي دائما وإلهامي. عائلتي الكريمة، لم أكن لأستطيع ان احقق ما حققته بدونكم. وكما علمتنا يا أبي دائما آقلب واحد، فكر واحدْ. هذه الدكتوراه ليست لعمر فردا ولكم لنا جميعا عائلة في قلب وفكر واحد.

I am grateful to all my friends who have supported and encouraged me throughout my journey, especially those who have been there every step of the way. To my friends in Egypt, Ali Hafez, Islam Amr, Ziad Shahin, Omar Elhamshary, Mahmoud Gouda, and Hassan Ayman, thank you for covering for me in countless situations while I was away. I also want to extend my gratitude to my close friends in Italy, Nurmuhammed Karimov, Khalid Helal, Ahmed Hamed, Abdelrahman Hossam, and Petra Marzova, for their unwavering support and encouragement.

Le mie più sentite grazie vanno ai miei supervisori, Francesco Rea e Giulio Sandini, per avermi dato l'opportunità di svolgere il mio dottorato sotto la loro guida. È stata un'esperienza incredibile e sono molto grato per il loro supporto. Inoltre, desidero ringraziare Alessandra Sciutti per il suo supporto professionale e le discussioni ispiratrici.

A special thanks to Vishwanathan Mohan, who has been a constant source of support since my master's degree. He supervised me during my abroad period at the University of Essex and has been a great friend and inspiration. I would also like to express my appreciation to Matthew Tata, and Lukas Grasse for having great collaborations throughout the research journey and in my period at the University of Lethbridge, Canada. In addition to Lukas and Matt, thanks to Claudia Gonzalez, Nicole Van Rootselaarfor, Mohamed Hafez, and Rezk Sedky for making my time in Canada productive, entertaining, and fruitful. Thanks to Doreen Jirak for having great collaborative work. I truly appreciate the opportunity to work alongside a supportive colleague.

Ai miei colleghi nei laboratori RBCS e CONTACT, desidero ringraziarvi per il vostro supporto e incoraggiamento. In particolare, vorrei ringraziare il fantastico gruppo della mia annata di dottorato: Luca Garello, Carlo Mazzola, Linda Lastrico, Giulia Albanese e Motonobu Aoki. Vorrei anche ringraziare il mio gruppo di arrampicata, Sara Mongile, Vasco Fanti, Matti Pava e Francesca Cocchella.

Infine, vorrei esprimere la mia gratitudine a Michele Tirone e Annalisa Scaletta per il loro aiuto con la lingua italiana. Nonostante i loro impegni, hanno dedicato molto tempo a insegnarmi, permettendomi di comunicare meglio e integrarmi più facilmente nella società. Grazie a tutti per il vostro sostegno e incoraggiamento; non avrei potuto farcela senza di voi.

Last but not least, I would like to thank the family of Egypt Scholars Inc.. Volunteering alongside great scholars has given me a lot of inspiration and also helped me to have a wider understanding of the scholars' communities.

Abstract

For many years, robots have been used in specific repetitive tasks, especially in industrial contexts. However, in recent years, robots start to be deployed in interactive and collaborative contexts with humans. The cognitive capabilities of robots are one of the main open challenges for effective interactions. Shared Perception is one of the important skills that are important for effective collaboration. In robotics, shared perception was studied from the human perspective (how to enable shared perception in an interaction with a robot). In the cognitive architectures side of research, shared perception was never studied except for some skills that are important to enable shared perception (perspective talking, gaze understanding, and gaze following). Therefore, my research firstly bases five general required skills for robotics in shared perception which are Having a common representation, Expressing effective communication, Spatiotemporal coordination, Affective modulation mechanism, and Understanding the other. Indeed, it is a complex skill and requires more than a Ph.D. to cover all concepts. Therefore, the main research activities were building cognitive architectures that try to address different concepts within the first three skills. The main aim is to build cognitive architectures that take the robots one step towards shared perception cognitive architecture. The architectures are built sequentially and based on each other. The approach to building these architectures has four characteristics Biological inspiration, Multi-modality more specifically audio and vision, Generalization (Not targeting a specific task), and Attention-based (Starting with state-of-the-art attention models and building upwards to include higher cognitive capabilities). Following this, the Ph.D. has three main research questions as the following:

1. How can we integrate state-of-the-art vision and audio models to allow the robot to jointly attend to the environment with a human partner? Is the behavior of the robot effectively received by the human partner? and What is the mutual influence between the robot and the human partner during the interaction using this architecture?

2. How can this integrated audio-visual attention architecture be used by the robot to understand a complex audio-visual environment? How can uncertainty be handled? How can the robot actively perceive the environment?

3. Can this perception architecture be generalized to different robots and applied to a complex task that requires coordination with another agent?

Each question of these three questions is related to one or more skills within the first three required skills from the mentioned five skills above. Trying to address these questions, I designed a series of architectures that are implemented cumulatively and showed how different cognitive blocks can be integrated to improve the perception capabilities of the robot, dealing with uncertainty, and noise in general conditions. The validation of these architectures was done using multiple robotic platforms (iCub, Pepper robot, and Essex agricultural robot) in different conditions with the robot only, with the existence of a human partner, with another robotic agent, and in a real-world application.

Further, I did some research activities related to improving the auditory modality. This was due to the outcomes of the experiments where audio processing was the main bottleneck process in the system. The improvements were building a developmental pipeline for audio. I used this pipeline to create an alternative model for audio localization that achieved very promising results. The second activity in auditory improvements was exploring alternative learning processes other than deep learning models that are more lightweight and suitable for robotic applications. Although the proposed architectures don't address all the required skills of shared perception, they address some points in that direction and also some open challenges in the field of cognitive architecture. More specifically, mutual influence in human-robot interaction scenarios, cross-modal interaction, and general unified perception modeling. The last integrated proposed architecture is a solid base for the development of a shared perception cognitive architecture for robots. Finally, this work opens multiple research lines in the future. The future development research lines can be divided into three main categories which are Integrating other cognitive components, Improving the individual modalities, and Applications, and Modeling the impairments.

Key-words: Cognitive Architectures, Shared Perception, Multisensory Integration, Cross-Modal Interaction, Biological Inspiration.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Rational discussion is useful only when there
is a significant base of shared assumptions.

Noam Chomsky

1.1 Motivation

Robotics showed a great impact on human life in different ways. In industrial settings, robotics are now deployed in many activities to advance automation such as packing, assembly, inspection, painting, and welding. This use of robotics improved industrial productivity increased the quality level and reduced the costs of production. In healthcare applications, robotics showed advancements in rehabilitation robotic devices, robotic prosthetics, and assisting in surgeries. This shows the great positive impact of robotic technologies on humans.

Reaching these success levels of robotics applications is also attributed to the improvements of the sensory and motor technologies that have reached (or almost reached) the required necessary levels including accuracy and speed. One other aspect of robotics is cognition. Cognition in robotics is the set of skills and processing capabilities that empower the robot with the requirements of understanding and interaction in the surrounding world. It is a multidisciplinary field of research that involves cognitive science, psychology, neuroscience, and artificial intelligence. The development in the cognitive organization of robotics is still emerging and in the early stages. The limitation in the cognitive robotic capabilities is the main reason why the deployed applications of robotics in the real world are mainly in a setting where the environment is controlled, the task is repetitive, and the social interaction with humans is minimal [217].

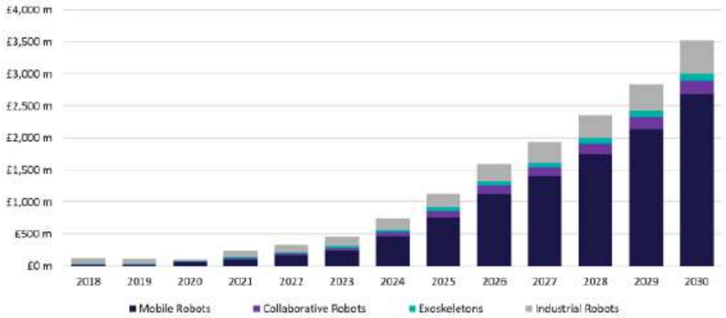


Figure 1.1 The robot market growth estimation by 2030 in the UK [84]

The cognitive development of robotics is aiming to endow the robots with human-like cognition. A robot with a human-like cognition is expected to understand human partners during an interaction and respond to human behavior in a way that is natural for humans. This will lead to more effective and safe interactions as well as solving more complex cognitive tasks. Additionally, human-like cognition is enabling robots to be socially integrated into human societies, unlike passive machine behaviors. In the meantime, the need for interactive robots with better cognitive capabilities is increasing especially in human-robot interaction scenarios, including social interactions, collaborative tasks, and autonomy in shared environments.

The challenges in the development of cognitive robotics exist in different cognitive capabilities such as perception, learning, action planning, natural language understanding and processing, and dealing with uncertainty. The development of robotic cognition is partially lagging behind the recent developments in cognitive science, neuroscience, and behavioral psychology. However, Robots can be also used as a tool to understand human cognition. For this reason, the research in cognitive robotics is important for creating effective collaborative agents that can support humans and also help us to understand human cognitive capabilities and behaviors. The research in cognitive robotics also tries to address explainability. Unlike the end-to-end models (Classical machine and deep learning models) that solve complex problems but do not provide reasoning for the decisions and actions taken by these models.

A recent report on the robotics impact in the UK [84] showed the expected growth of robotics in different sectors until 2030 (as shown in the figure 1.1). The report also highlights the challenge of acting and performing in unstructured and hard environments. Thus, the focus on cognitive architectures that endow the robots with better cognitive skills is an important requirement that will strongly address this challenge.

Accordingly, building architectures that achieve human-robot shared experiences is one of the main research focuses in the meantime [65]. Endowing the robot with a shared experience with a human partner will have an impact on human-robot social collaborations and interactions. On this premise, modeling shared perception on robots is important for the goal of having social robotic collaboration. Shared perception is the perception of the shared environment/stimulus by two (or more) observers simultaneously which also includes the effect of the partner's implicit and explicit actions on the observed stimulus/element in the environment. Shared perception is a crucial skill during an interaction that enables safe, efficient, and effective collaborative work in the shared environment. In the literature, shared perception is addressed from the human side (how a robot affects the perception of the human is a shared scenario), or partially from the robot side by addressing some skills such as perspective taking and gaze understanding and gaze following. But looking from a broader point of view, enabling human-robot shared perception for both agents (not only the human) requires the robot to have a set of skills. We categorize these skills as the following:

- **Having a Common Representation**: Sharing the same understanding of the human partner about the environment as well as objects and agents within the environment. This skill also includes the capability of dealing with uncertain information due to existing noise, the lack of redundancy from different modalities in some cases, and inconsistencies in the perceived sensory information. Having a common representation engages different levels of commonality from a basic common understanding of feature space to a much higher level of commonality that shares a common affective representation.

- **Expressing Effective Communication** Effective implicit and explicit communication through different sensory channels with the other agents in the environment. This includes evoking the communication cues within the shared environment that can be interpreted by the human partner as meaningful social communicative cues.

- **Spatiotemporal Coordination** Perceiving, acting in the environment and reacting to the cues from other agents in the environment in real-time consistently. It enables active interaction and needs high-level planning and execution of an agreed policy between the agents.

- **Affective Modulation Mechanism** Affective modulation is a core element in shared perception. It is the influence of the affective state of others on the observer's perception in the shared context [221]. Robots need to be capable of expressing an affective state.

Additionally, to achieve a human-like shared perception, the perception state of the robot should be influenced by the affective state of the human partner. This includes the capability of expressing emotions and other affective states (i.e. trust, comfortability,.. etc.) and understanding the other agent's (human) states.

- **Understanding Others** Interpreting the implicit and explicit actions of the other agent (human). This includes social cues (i.e. Gaze, facial expressions, ...etc.), perspective taking, and intentions.

These skills are strongly interconnected, linked together, and need to be studied and modeled on robots. Additionally, it is required for all these skills to engage multimodality in their processes. Interpreting, coordinating, and integrating information from different modalities (i.e. vision, audio, tactile, ..etc) enables building a more complete and accurate representation of the surroundings. On the other hand, acting using multimodality also increases the kinds of interactions and the cues that the robot can execute.

1.2 Aim and Approach

My research focuses on building different cognitive architectures for robots in the direction of having a human-like shared perception cognitive architecture for robots. These cognitive architectures have four main characteristics as the following:

- **Biological Inspiration**. Getting inspired by neuroscience, cognitive science, and behavioral psychology in the implementation of the architecture.
- **Multimodality**. Or more specifically, audio and vision.
- **Generalized settings**. A general approach that can be deployed in different robotic platforms, applied in different conditions and applications, and task-irrelevant to insure expandability and modularity.
- **Attention Based**. Modeling the human process by starting with attention and moving toward higher cognitive processes.

Indeed, the focus of the research is on a set of skills and not addressing all aspects of shared perception. However, the aim is to build a scalable base starting from existing models of attention towards a unified architecture for multimodal shared perception for robots.

1.3 Research Questions

Following the main aim of my research, I divided the research into three main research questions that have been formulated as the following:

RQ1: How can we integrate state-of-the-art vision and audio models to allow the robot to jointly attend to the environment with a human partner? Is the behavior of the robot effectively received by the human partner? and What is the mutual influence between the robot and the human partner during the interaction using this architecture?

RQ2: How can this integrated audio-visual attention architecture be used by the robot to understand a complex audio-visual environment? How can uncertainty be handled? How can the robot actively perceive the environment?

RQ3: Can this perception architecture be generalized to different robots and applied to a complex task that requires coordination with another agent?

[*RQ1:*] is aiming to take a step toward having a common attentional representation of the environment (the robot and a human partner jointly attend to the stimulus). Additionally, it examines the efficiency of the robot's expressions and how they influence the attention of the human partner. Finally, it deals with how the existence of human partners and their actions in the environment affects the attentional behavior of the robot.

[*RQ2:*] is focusing more on building a common representation of the environment by adopting a bioinspired modeling approach. Additionally, it addresses how motor actions can be engaged to improve the perception processes (endowing the robot with proactive behavior in perception like humans).

[*RQ3:*] is addressing the action coordination in the spatiotemporal domain with other agents. Additionally, it addresses how the architecture can be applied in different robotic platforms and how the architecture can be generalized in a real-world task.

Figure 1.2 shows the connection between the research questions and the relative shared perception required skills. It shows the first three categories (which have the main focus of this thesis) and some of the skills within the category. Also marks the relevance of the research questions with the corresponding skill. It is also important to note that the used graphical images for the skill category are trying to give a visual representation of the category which will be used in this thesis in other figures.

In this thesis, I explain how the research is addressing these research questions using the explained approach (four characteristics). Additionally, it tries to address the limitations that

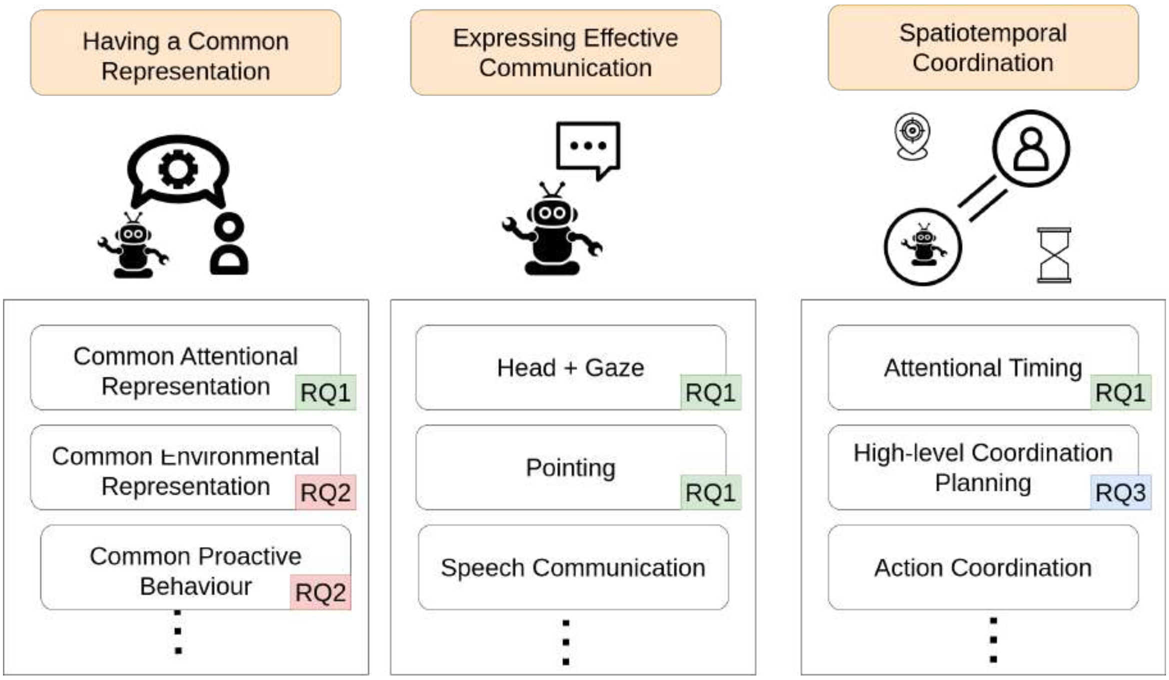

Having a Common Representation
Expressing Effective Communication
Spatiotemporal Coordination
Common Attentional Representation
RQ1
Common Environmental Representation
RQ2
Common Proactive Behaviour
RQ2
Head + Gaze
RQ1
Pointing
RQ1
Speech Communication
Attentional Timing
RQ1
High-level Coordination Planning
RQ3
Action Coordination

Figure 1.2 Research Questions and Shared Perception Skills Connections

were faced during the thesis. These limitations are generally in auditory processing and a dedicated part of this thesis is explaining the approaches to tackle these limitations.

1.4 Thesis Structure

As shown in Figure 1.3, the thesis comprises of five parts. For each chapter within the part, the figure also indicates the main contribution of this chapter. Additionally, the figure also connects the chapters with the research questions and the shared perception category (by using its graphical representation besides the contribution of the chapter/part). This section will briefly talk about these five parts as the following:

Part I **"Background" :** is for giving a general background and the state of the art in **Chapter 2**. It goes through shared perception in robotics, cognitive architectures, and each of the four characteristics of the thesis approach in detail.

Part II **"Audio-Visual Cognitive Architecture For Joint Attention" :**This part of the thesis describes the development of a biologically inspired integrated audio-visual attention architecture, the behavior of the robot in a joint task with a human partner, and the mutual influence between them. This is mainly to address [*RQ1*]. This is done

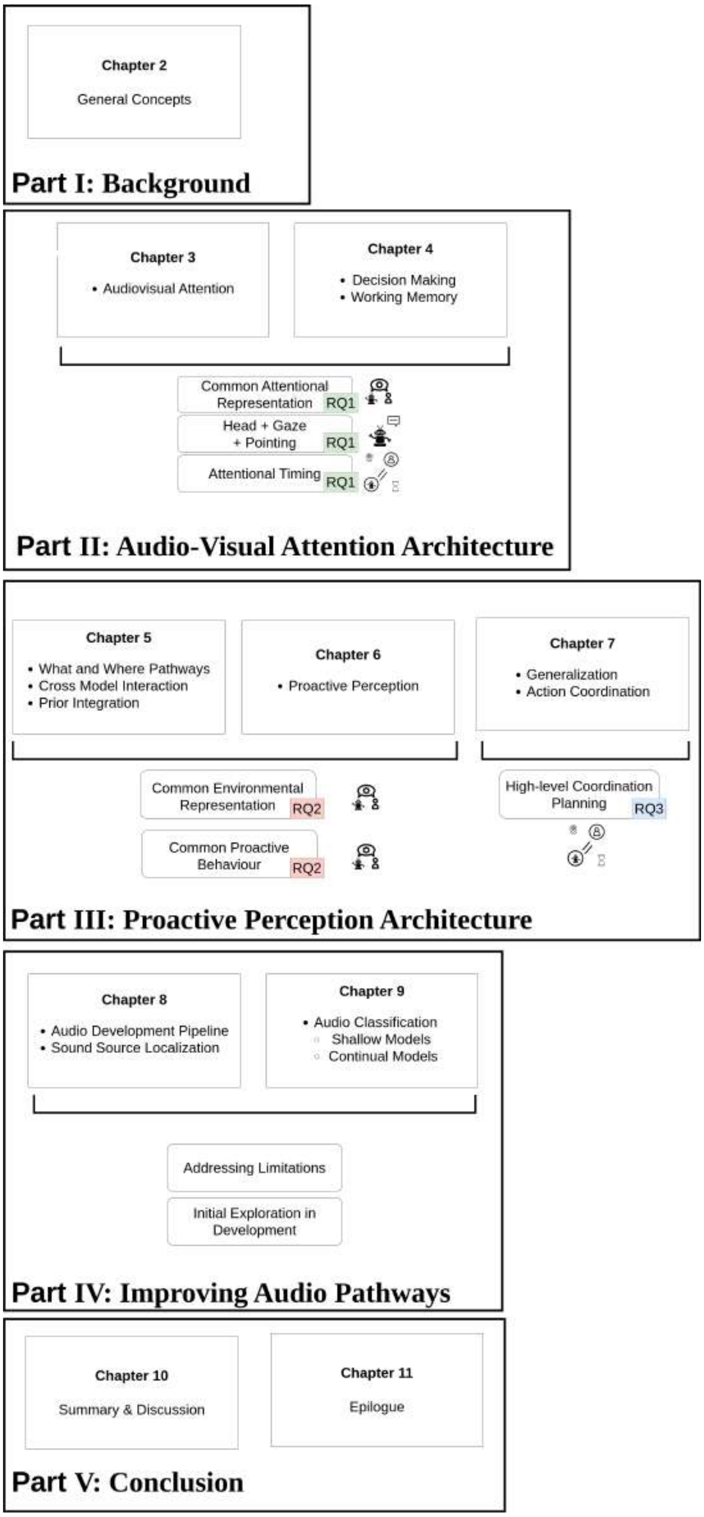

Chapter 2
General Concepts
Part I: Background
Chapter 3
• Audiovisual Attention
Chapter 4
• Decision Making
• Working Memory
Common Attentional Representation RQ1
Head + Gaze + Pointing RQ1
Attentional Timing RQ1
Part II: Audio-Visual Attention Architecture
Chapter 5
• What and Where Pathways
• Cross Model Interaction
• Prior Integration
Chapter 6
• Proactive Perception
Chapter 7
• Generalization
• Action Coordination
Common Environmental Representation RQ2
Common Proactive Behaviour RQ2
High-level Coordination Planning RQ3
Part III: Proactive Perception Architecture
Chapter 8
• Audio Development Pipeline
• Sound Source Localization
Chapter 9
• Audio Classification
◦ Shallow Models
◦ Continual Models
Addressing Limitations
Initial Exploration in Development
Part IV: Improving Audio Pathways
Chapter 10
Summary & Discussion
Chapter 11
Epilogue
Part V: Conclusion

Figure 1.3 Thesis Structure. Showing the five parts of the thesis and the different chapters with their connections to the Shared perception skills and the research questions.

through two versions of the architecture. **Chapter 3** is presenting the first version a preliminary study. In this study, the main contribution is in the attention model integration. while **Chapter 4** is the extended version of the study using a modified version of the attention architecture (Version 2) and extending it to include other cognitive components (Working Memory, and Decision Making).

Part III **"Proactive Audio-Visual Cognitive Architecture For Perception"** is addressing the general perception audio-visual model and addresses both *RQ2* and *RQ3*. It consists of three chapters. **Chapter 5** explain in detail the main architecture, how the architecture is structured using what and where pathways, how it addresses cross-modal interaction, and how it uses the previous associated experiences in perceiving the environment. **Chapter 6** is exploring how can actions be included in the architecture to improve the perception focusing on the effect on a sub-part of the architecture which is sound localization. The final Chapter in this part is **Chapter 7** is addressing the generalization of the architecture and the implementation of the use case in a real-world coordination task.

Generally, [*Part II*] and [*Part III*] are presenting different versions of the architecture which represent the development journey. Each version of the architecture is presented in a separate chapter starting by **Chapter 3** to **Chapter 7**. Each chapter goes throw the biological inspiration, the implementation of the architecture, the executed experiment to examine this architecture, and finally the discussion and conclusions that can be drawn from the results of the experiment.

Part IV **"Improving Audio Pathways: Alternative Pipelines For Developing Robot's Audition" :** This Part aims to address the limitations that were shown in the behavior of the robot using the developed architectures in the previous parts (Part Two, and Three). These limitations are generally related to auditory processing due to the complexity of the audio signals and the noise. Therefore, this point was addressed by proposing different developmental pipelines for robot audition. This part is divided into two chapters. (**Chapter 8** is proposing a novel pipeline for sound source localization. and (**Chapter 9** is exploring alternative learning techniques in for auditory classification.

Part V **"Conclusion":** It has the summary and discussion chapter (**Chapter 10**) where an overview and discussion of the outcomes of the thesis, novelty and future work will be drawn. Finally the Epilogue of the thesis (**Chapter 11**).

Part I

Background

Chapter 2

General Concepts

> If I have seen further than others, it is by
> standing upon the shoulders of giants.
>
> **Isaac Newton**

In the first chapter, I explained the main motivation, the aim, the research approach, and the research questions. In this chapter, I will explain in detail these points and the state-of-the-art regarding them. Firstly, I will go through the motivation (Shared perception in robotics) then I will explain in brief some concepts related to cognitive architectures and their components that are relevant to my research work. Further, I will go through each of the four points of my approach (Biological Inspiration, Multi-Modality, Generalization, and Attention Based). Finally, I will summarize this chapter as sort of positioning my research work in relation to the research community.

2.1 Shared Perception in Robotics

With the increasing role of robotics in human life, having effective socially collaborative robots has become an important requirement. In collaborative human-robot interaction scenarios, the robot and the human partner/s are sharing together the same environment. Perception is one of the important components in effective collaborations. Perception is the cognitive process that an agent performs by processing the information from the scene and understanding it. In the context of multiple agents observing the same environment simultaneously, the perception is shared in the social context between these agents. It is shared because of the social effect on the way one agent perceives the shared cue in the scene (with the other social agent) [145].

2.1.1 Integrating other's attention, perspective, actions, and inner states

Shared perception is a new term in robotics. It is framed based on how biological creatures are integrating the other's cognitive process, more specifically, attention, perspective, and actions. The following paragraphs will explain these components in the context of shared perception.

Attention Staring with integrating other's attention, gaze following and joint attention are two of the cognitive skills that can be represented as an integration of other's attention. Butterworth has defined gaze following as "looking where someone else is looking" [31]. when looking at joint attention, the skill requires a higher level of cognitive processing than gaze following as it is a triadic interaction between two agents. Joint attention skills are developed in infants between 9 and 15 months [238].

Perspective Moving forward, perspective talking is what we refer to here with integrating other's perspectives. In this research field, many pieces of research had been conducted on different aspects including modeling perspective talking in robotics [58], using robots to help children in perceiving the world from its perspective [126], and studying the effects of perspective taking in a human-robot interaction scenario [144].

Actions The actions of others may create an effect on one's perception. More specifically, humans are capable of understanding implicit cues within the action that leads to the perception of some information. For example, Kaiser et al. showed that children can estimate the weight of an object from the observations of videos [104]. These observations further improve the accuracy of perceptual judgments [73]. These studies influenced the robotics community to study the effect of robotic actions on the perception of the human observer [130, 222].

Inner States One way of understanding the behavior of others is mentalization. It is understanding the other's inner states including needs, thoughts, feelings, goals, and reasons. It is also common in literature by the term "Theory of Mind" [193]. This social cognitive skill is been hypothesized to be developed starting from infancy [149] and gradually over years[237]. Recent studies represent the inferences of other's inner state as a top-down modulation process for perception [12]. In fact, research showed that implicit information from the action and explicit knowledge about the intention of the person who does the action leads to modulation of the action-perception [91]. Also, other studies showed a modulation in following the gaze [233] and the cycle of perception and action [137]. To summarize, others' inner states have an effect on the perception of the environment as well as the collaboration with them.

In robotics, many studies have been conducted to examine the effect of the robot's inner states and showed that the inner states of the robot are affecting the human perception and action processes (check [234] for a review).

We demonstrated here, how perception is integrating others' attention, perspective, actions, and inner states. This integration helps infants to develop their cognitive skills (both social and non-social skills), leads to the creation of common ground, and improves collaboration mechanisms. Additionally, studies showed that robots can invoke this shared perception behavior with their human partner.

2.1.2 Shared Perception From the Robot's Side

Currently, as shown in the previous section, in the shared perception context, researchers are typically addressing the human side or a specific skill that is required to be enabled in robotics to accomplish a shared perception scenario for the human side. However, some topics related to shared perception were studied separately (Not in the context of shared perception) for the robot's side such as perspective taking, gaze understanding, and gaze following. Modeling the shared perception processes in the robot would import the shown advantages of this behavior such as developing cognitive skills and effective collaborations. But this is a challenging task especially because the studies conducted to understand this behavior on the human side are bypassing a lot of primary skills that are currently missing in the robots. For example, the spatiotemporal coordination between humans that shared perception occurs in between. Both humans have coordination in time and space (there is almost no lag of perception if they both are looking to the same object). While in robotics, robots might need a longer time to process some information in comparison to a human partner.

Therefore, we first framed the general development directions or skill categories for the robot side to enable shared perception for both the robot and the human (mutual influence). We briefly presented them in the motivational part (section 1.1) which are:

- **Having a Common Representation** Recent studies showed that the robot has to find the same stimuli salient and have similar ways of using their sensory systems to promote mutual behavior [25]. According to this, we suggest that to reach an effective level of shared perception, the robot has to have a human-like representation. The idea in the cognitive architecture domain is commonly known as a biologically inspired approach (which will be discussed in section 2.3). Thus, using human-like models in attention, perception, and action is a key point. It is also important to point that there are multiple levels of commonality, a basic minimal level can be the representation of the scene's feature in space and time. and a higher level representation can be sharing the affective states. Indeed, it is a circular process, which means that human-like modeling increase the shared representation, and the shared representation leads to sharing more commonalities (creates a common ground).

- **Expressing Effective Communication** Empowering the robot with the capabilities of expressing communicative signals that can be understood correctly by the human partner. These signals can be explicit or implicit. This is a crucial aspect to enable shared perception on the human side. This importance is shown in the study which elaborated on the differences between the mechanical robot and social robot in enabling shared perception [145]. Currently, robots have the capability of expressing such cues. However, the execution of these cues has to be enabled actively based on the perception of the robot. Unlike the currently applied methodologies in the experiment where the robot is programmed to act in a passive way.

- **Spatiotemporal Coordination** Coordination in time and space is a very important point in any shared task. So when it comes to sharing the perceptual events and states, the coordination has to be in perceiving the event and responding to it. This includes the low-level attentional response time and up to high-level action planning. This coordination has to have an agreed policy to occur. This policy can be part of the perceptual representation in a bottom-up fashion and explicitly agreed on as a top-down property.

- **Affective Modulation Mechanism** In this point, we point to the skill of communicating affective cues that are capable of modulating others' emotions. In fact, this skill can be categorized under the communication point. However, we categorize it to highlight the importance of the affective aspect. The importance of this skill in shared perception started to get more attention in recent years. In fact, a recent special issue was published to highlight this aspect [221].

- **Understanding Others** Encoding the implicit and explicit cues of others through processing their attention, perspective, actions, and inner states. Further, building a model about their goals and intentions through this information. In this category, some advancements had been made such as in perspective-taking, and gaze understanding. However, these skills have to be integrated together and encoded through prediction

We mentioned the skill categories that are required to reach a mutually shared perception between the robot and a human partner. To summarize, the current use of the current robotic perception models that are designed for better human-robot collaboration is limited to high-level behaviors. This does not allow the robot to have organized low-level sensorimotor modalities [2]. Accordingly, adopting a human-like perception in robotics would be an automatic way of understanding the human partner [59]. In fact, it has been suggested that robots have to exhibit similar attentional characteristics [26] and use sensories to be intuitive [25]. A recent study also showed the importance of shared perception in collaborative human-robot interaction [144]. The study showed that better performance was achieved when the robot adopt a shared perception strategy for the decision-making of suggestions to the human collaborative partner. Another study showed that when the robot and a human partner take mutual adaptive decisions the performance of the human-robot team increases [173]. Thus we take this as the motivation for our research and the general theme of the thesis. For each part of the thesis, I will be reviewing the recent work in more specific domains of cognitive architectures and robotics.

2.2 Cognitive Architecture

Cognitive architecture refers to the structural organization of the process involved in human cognition. It provides a comprehensive description of the functionalities and their interaction. Cognitive architectures often have a principal representation and a computational implementation that show how the information is processed, stored, and transformed in the mind. We suggest reading the review by Kotseruba and Tsotsos for the cognitive architectures that were developed last 40 years ago for more comprehension [120].

2.2.1 Cognitive Architectures Core Functionalities

The components of cognitive architecture can vary from one architecture to another based on the main focus of the architecture. These components are generally the core components of human cognition.

Attention Attention is a cognitive process that allows the agent to focus on a discrete amount of information in the scene while ignoring the rest of the perceived data. Attention has been defined as a selection process or as an allocation of limited resources. Attention as a selective process is referring to the agent's skill of actively selecting stimuli or information to process while ignoring the rest. Selective attention processing emphasizes the importance of the active role of the agent to influence the attentional process (the top-down process) as well as the bottom-up environmental salience-based influence. On the other hand, the other idea about attention as an allocation of limited resources describes attention as limited processing cognitive capacity. It is based on the view of limited processing capabilities an agent can perform at an instant of time. Although attention has been studied extensively in different research domains, It is difficult to draw a general definition for it. Therefore, different studies focused on different kinds of attention (i.e. selective attention, divided attention, executive attention, etc.).

Perception is the process of organizing and interpreting the sensed information from different sensor modalities (i.e. Visual, Auditory, and Haptic). The perception process allows the agent to understand the environment including objects, people, events, and actions.

Action and action selection are the proactive process that changes in the state (typically the physical). Cognition studies usually focus on planning, selecting, and executing the action. Additionally, the direct and proactive role of actions on other cognitive processing such as perception, memory, and learning.

Memory is the cognitive component where the information is stored. The stored information or knowledge can vary in types and functionalities such as short-term memory, working memory, long-term memory, and associated memory.

2.2.2 The Three Types of Cognitive Architectures

In the last decade, many cognitive architectures have been proposed. A recent common categorization method for these architectures is based on the kind of represented and processed information in these architectures. It divides the architectures into three major taxonomies. the symbolic architectures (cognitivist), the emergent architectures, and finally the hybrid approach architectures. The symbolic (cognitivist) approach is the classical one which is using rules to manipulate symbolic representations of the concepts. This approach embraces the view of cognition as a kind of symbolic computation. On the other hand, the emergent approach is more into real-world temporal processing using connected, dynamic, and enactive methodologies. It is based on the view of cognition as an organizational process in different nodes. Due to the limitations of each of these two approaches (both of them aren't capable of addressing all major processes of cognition), the hybrid approach tries to combine the advantages of both the symbolic and emergent approaches together.

2.2.3 Cognitive Robotics

In the robotics field, the focus of development on the software side has different directions. In 2019, Murphy referred to the approach that focuses on developing tools and blocks using the models of traditional artificial intelligence as "intelligent robotics". However, the cognitive robotics field is not the same as intelligent robotics. The cognitive robotics term originated in the 1990s. From that time on, multiple definitions have been proposed [230, 28, 44, 133]. The most recent definition that we believe that it gives an excellent comprehensive definition of cognitive robotics is the one proposed by Cangelosi and Asada is: "Cognitive Robotics is the field that combines insights and methods from AI, as well as cognitive and biological sciences, to robotics." [32]. So, It goes beyond AI and includes interdisciplinary fields. Going more in practical terms, cognitive robotics is a sub-domain of robotics research focusing on how to endow a robotic agent with cognitive capabilities such as perception, learning, and reasoning. It deals with the robots as embodied agents in the environment, more specifically in real-time interactions.

In recent years, cognitive robotics has received the attention of researchers under many research projects funded by financial institutions including DARPA, the European Commission, the UK government's office of science and technology, Japan Science and Technology Agency Exploratory Research, and others. This actually shows the importance and potential of this field.

One of the uses of cognitive architectures is the examination of modeling human cognition and behavior including brain activities [232]. Following this concept, cognitive robots can be used to validate the proposed architectures and implementations. The advantage of robotics over the classical computational approaches and artificial systems is that robots provide realistic and time-oriented interactions with the environment. Additionally, the environment can include other social and non-social agents during the interaction.

In the research of cognitive robotics, there are different branches of robotics in the community. They can be divided based on their cognitive functionality (i.e. cognitive perception, cognitive navigation, cognitive manipulation, and social cognition). Another way of categorization is based on applied behavior. (i.e. swarm cognitive robots, soft cognitive

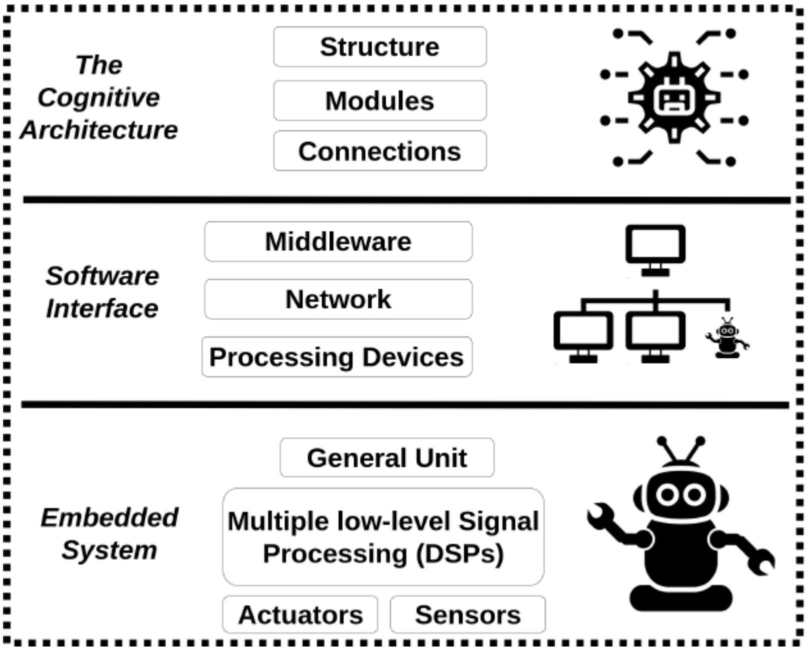


Figure 2.1 Cognitve Robotic System Layers

robots, developmental cognitive robots.) The software implementation of cognitive robotics has different layers which I will present in the following section.

Implementations layers of the cognitive robotics

The development of cognitive robotics generally consists of three layers shown in figure 2.1 The embedded mechatronics system, the middleware and processing units, and finally the cognitive architecture.

Embedded Layer The embedded layer is the lowest layer in processing. it deals with sensors and actuators. Also, it defines the embodiment of the robot based on the shape, fixation, and capabilities (degrees of freedom) of the design. In this layer, digital signal processing units are usually associated with the sensors and actuators. The highest part of this layer is a generic unit that communicates with the upper layer. This general unit is the main low-level processor of the robot.

Middleware The middleware layer is responsible for connecting different processing units altogether including the robot. The processing units can be other PCs or GPUs in the network.

Middlewares are designed to offer communication channels with different devices, different operating systems (Linux, Windows, macOS), and different software languages (Python, C++, Java, ..etc.).

YARP (Yet Another Robot Platform) [150] is the developed middleware for the iCub robot. Alongside YARP, ROS (Robot Operating System) [195] is a widely used middleware in different robots. One of YARP's advantages is that it has a bridge to operate alongside ROS. So, different modules developed on top of ROS can easily communicate with other modules in YARP and vise-versa.

The cognitive architecture The cognitive architecture layer is the top layer of the whole system. It is built using the middleware of the system. This layer consists of the different modules (components) of the architecture. Each module is operating as a separate entity that communicates with other modules through the connections. The connections generally send and receive data or commands.

2.2.4 Examples of Cognitive Architectures

In recent years, a lot of cognitive models have been proposed for robots. In this section, we will mention a brief about these models based on cognitive processes.

Among the cognitive architectures, CLARION [231], SOAR [125], and ACT-R [6] are one of the common architectures. The most presented modality in the literature is vision. Most of the proposed models are operating with very specified scenarios/tasks and generally use computer vision methodologies and toolkits. Only a few models such as ART [179], SASE [252], Leabra [179], and Darwin [223] are biologically inspired and are designed to model a specific process and limited applications in control environments.

On the other hand, auditory perception models are less common in cognitive robotics research. Most of the focus in the auditory models for robots was on the linguistic information carried in the audio in the context of speech commands to the robotic agent [120] such as CORTEX [210] and Ymir [235]. Some other models have an auditory perception component to study the multimodal integration or for specific tasks such as localizing speakers in the scene [120, 72].

2.3 Biologically Inspired Cognitive Architectures

The term Biologically Inspired Cognitive Architecture (BICA) started to rise through a program funded by DARPA in 2005. Further, the term was adopted in different research communities [69]. In general, BICA is designed to show and simulate similar functionalities of the biological brain, based on structuring the cognitive process and their connectivity. BICAs get inspiration from the brain for representing knowledge and, unlike the subsection of the cognitive architectures that are modeling the mind process almost exclusively [122]. It was also suggested that biological inspiration is one of the keys to the next generation of artificial cognitive agents [38].

The gap that BICAs are trying to fulfill is generally in three main domains as the following:

- An approach to understanding intelligence. Modeling complex integrated behaviors can examine existing ideas about cognitive processing. Consequently, a better understanding of these processes and hopefully addressing new solutions for impairment and diseases.

- Developing embodied machines (Robots) that behave like a human. This point is to address the physical and social interaction between the robot, environment, and other agents in a natural way.

- The efficiency of using the sensory and computational resources. BICAs are generally trying to minimize the resources such as using binaural hearing instead of solving the problems by increasing the dimensionality of the sensories/processing power for example using a microphone matrix instead of dual microphones for sensorial processing and using a huge dataset in learning processes.

The main objective was to develop artificial models for the human-intelligence. Using these models, artificial machines are equipped with the required intelligence to solve complex tasks in complex environments. Thus, the machines are also capable to adopt and learn effectively with autonomy [216]. In the meantime, there are several challenges associated with the BICAs. The complexity of biological intelligence makes modeling it a challenging task. This is one of the primary challenges for BICAs. Another challenge is the requirement for interdisciplinary collaboration within the research community. This includes psychology, neuroscience, engineering, and cognitive science. The collaboration will foster development which can be seen in some recent research.

One of the advantages of having human-like computational behavioral models is improving the efficiency of the task in the human-robot interaction scenario. For example, it was shown that a human-like gaze improved the timing of a handover task between a robot and a human [157]. Another research showed that the human-like gazing and pointing behavior has enabled the effect of shared perception and positively improved the accuracy and reduced the perceptual errors in a shared task [145].

Finally, we would like to suggest the review by Samsonvich for more informative details about the biologically inspired cognitive architecture [216]. It is not a very recent review. However, it elaborates on the challenges and the road map of this field.

2.4 Multimodality

Humans are receiving a stream of information from multiple sensory modalities (i.e. Vision, audio, touch, .. etc.). The brain receives this information and integrates them coherently to create a unified representation of the environment. The integration of the information coming from different sensors is usually called sensor data fusion in robotics [110] and features integration in cognitive science [260].

Perceiving the world using multiple senses gives more information about the environment which creates a higher level of situational awareness. Multimodal perception improves the robustness and accuracy of perception. Additionally, it increases confidence and decreases ambiguity. This is due to the capability of integrating information from multiple sources and not relying only on a single source. Indeed the information coming from different modalities has some redundancy as well as complementary parts. Therefore, having a multimodal artificial system is a big advantage that increases the range of applications where this system can be deployed and used. However, integrating the information (sensor fusion, or feature integration) is a complex process which is usually a challenge. These challenges will be discussed further in this section.

Many proposed cognitive architectures are having two or more sensory modalities. The integration methodology varies usually based on the level of abstraction in the architecture. However, none of these architectures addressed cross-modal interaction. Cross-modal interaction is a widely agreed phenomenon with evidence from neuroscience and psychology. Basically, it is the effect of one sensory modality on altering another sensory. Altering the senses induces different effects (i.e. McGurk effect [143], Ventriloquist effect [41], Double-flash illusion[225, 224]). These effects actually give us more information about the biological cross-modal integration but to the best of our literature research, cross-modal interaction is not addressed in the cognitive architecture domain [120].

In summary, there are some challenges in multi-sensory integration in cognitive architecture as the following:

- **Handling data-related fusion aspects**. Multi-sensory integration generally (not only in the domain of cognitive architecture), has a challenge on the methodologies of handling the data fusion-related challenges which includes uncertainty, inconsistency, disparateness, and conflict situations.
- **Cross-model interaction**. As mentioned in a previous section, this domain is not yet addressed in any of the previously proposed architectures
- **Generalization**. Most of the proposed multimodal architectures that consider generalization among sensor modalities are very task specified. For example, in [138], the authors proposed a system for multimodal emotional intelligence that generalizes the understanding of emotions using different modalities. Although it performs a generalization on multiple modalities, it is a trained model for a specific task (Recognizing and expressing emotions). Generally, the architectures which are designed to perform sensor fusion are operating under specific conditions and to reach a specified goal. And there isn't a universal proposed solution in the robotics domain. This is also related to encoding the knowledge in generalized approaches with different varieties of formalization that can be transferred from one modality to another, and from one task to another [127].

2.5 Generalization

As robots are now engaged in more applications, the environments where the robot is operating are usually complex. Therefore, It is important for a cognitive robot to be able to generalize the processed information from the different senses. This includes handling the uncertainty of the perceived elements in the environment, behaviors of human partners, and also communicated goals. This has been highlighted in a very recent survey for cognitive robotics [226].

Trying to address this point and as mentioned, most of in multi-model cognitive architecture didn't address a unified technique to integrate the information from different senses, I put the generalization as a core point in the thesis approach. This means that the aim isn't towards solving a specified task in the robotics domain, but more towards a generic approach. Also, the approach while designing the architectures is to focus on handling the uncertainty which is generally a consequence of operating in complex and unstructured environments. Of course, when testing the architecture, it has to be within defined processes, but the goal of this point is to minimize the constraints, design architectures that can fit different tasks, address real-world scenarios, and use different robotic platforms. Both generalization and multi-modality are two key challenging points of the next-generation robotic platforms that will be interacting in diverse natural conditions [255].

2.6 Attention Architectures

Attention is an important component of cognition. Attention is a set of proactive processing mechanisms that are not only affecting perception but also other cognitive processes. There are different taxonomies of the attention mechanisms that are generally differentiated by the kind and amount of the suppressed information and information in the current attentional focus.

2.6.1 Visual Attention

Visual attention is the type of attention that has received the most focus. In cognitive architectures, attention as a selection process is the most common viewpoint. From the computational side of view, Itti and Koch proposed a bottom-up selective attention model. The model is computed as an integrated saliency attentional map [239] [212]. The integrated map is a linear combination of multiple low-level saliency features map such as (Edges, chrominance, and intensity). Based on this computational model, PROVISION (PROactive VISual attentION) architecture has been implemented [201] and used in robotics in multiple applications [247, 72]. The PROVISION model includes actions in the process of attention which included prediction, top-down control, and proactive action in the attentional process. This can also be considered an extension of the proposed model of Itti and Koch. The PROVISION model provides a modular tool for bottom-up visual attention enabling the ability to tune the importance of particular visual stimuli, for example, forcing the attention towards a bright object by putting more weight on the intensity value.

2.6.2 Auditory Attention

Although most of the focus was on visual attention, auditory attention is an important skill when dealing with the environment. The most common example of applications related to the use of this skill is the cocktail party problem [81]. Modeling auditory attention is more difficult than visual attention due to the complexity of the auditory signal which in most cases has some noise and multiple sources. Additionally, the validation process is more complex as there is no direct way to observe the attended cue as in vision (tracking the gaze). Auditory attention is based on multiple acoustic characteristics such as sound sources, rhythm, intensity, and harmonics). In robotics, auditory attention is usually modeled based on source localization (Spatial based attention). Check [108, 167] for a review.

The development of the auditory attention in this thesis is based on an existing bio-inspired Bayesian audio localization model [119]. The auditory attention component os built to redirect the attention of the robot toward salient auditory signals. The system is based on the biological basis of how humans perform sound localization. Humans use different cues to localize sound sources: the interaural time difference (ITD) and interaural level difference (ILD). Both are differently recruited by the auditory system to derive the direction of sound arrival. In our implementation, we focused on the ITD cue as the principal computational method since there is a robust literature that uses ITD for sound localization in artificial systems [8]. The general idea behind ITD is to infer the direction of a sound from the difference in time of arrival (TOA) between the two ears. Different approaches have been proposed in robotics to computing the TOA, the most common one is based on correlation metrics [89]. This approach performs well but is sensitive to noise and reverberation, which is problematic, especially in presence of ego noise produced by robots. Other biological systems in nature use ITD cues to localize sound by employing either bank of coincidence detectors connected by delay lines, as in the avian brainstem [98], or more complex phase-tuned mechanisms as in the mammalian brainstem [75]. The audio localization model used in this research modeled the spectral decomposition of the human basilar membrane with a gammatone filterbank and modeled delay-tuned units in the auditory pathway as banks of narrow-band delay-and-sum beamformers. To further deal with the spatial ambiguities associated with interaural cues [21], the model uses a Bayesian regression model that infers the location of the sound source using the previous results of the spatial localization values. As a result, the location is reliably estimated in the robot's allocentric coordinate frame as a probability distribution of sound source locations across azimuthal angles. This probability distribution is used to create an allocentric saliency map of the sound locations.

2.6.3 Audio-Visual Attention

Audio-visual attention is the combined attentional processing that is responsive for both modalities. The computational modeling of audio-visual attention in robotics is still lacking [64]. Only a few models have been proposed. For example, in [3] and [182], the authors modeled the audio-visual attention to explore the cross-model attention mechanism in humans. Some other models In robotics, the focus There are really few models that addressed audio-visual attention in robotics [64]. In these models [212, 128, 72] integrated the audio as a visual feature map in the linear combination process of visual attention. Yet, there is no evidence from the biological processing that this is the case. The focus of previously proposed audio-visual models in robotics focused mainly on the multisensory binding and for a specified task mostly speaker localization (e.g. [246, 198]) ignoring the low-level attentional selection and dynamics [64] It is clear that there is a gap in this direction and a need at the same time, especially in interaction scenarios where multiple agents are interacting with the robot.

2.7 Research Positioning

This thesis is a development base for shared perception cognitive architectures for robotics. This is through tackling some of the required skills for shared perception in robotics. The choice of skills was based on the complexity of the required processing. The starting point was the state-of-the-art models for audio attention and visual attention.

The PROVISION attention model [201] and the Bayesian audio localization model [200] are the starting point of the development of the thesis. A Bayesian localization model is an active approach that was developed on the iCub robot. The model is based on the interaural time difference (ITD) phenomenon. The model computes the posterior Bayesian probability and gives an output of a 360 degrees Bayesian map for the azimuth angle. The PROVISION model is an implementation of state-of-the-art biological attention methodologies. It computes a combined saliency map based on visual early features. My main research part is based on these two models. In my research, I took these two models further and developed different cognitive processes on top of them.

2.7.1 The Contribution

In this thesis, I present my research which contributes to the community with the following:

- In part I (Background) which is this chapter, I framed theoretically the requirement and research directions toward human-robot shared perception.
- In part II, my contribution is the design, implementation, and evaluation of an audio-visual cognitive architecture. The main novelty of the designed architecture is the biologically inspired time-variant decision-making process that addresses the speed-accuracy trade-off in the attentional decision. Additionally, we contribute with the evaluation process of this architecture as it considers a point that has been ignored in other audio-visual attention systems which is the mutual influence between the robot and the human in the attentional task that they jointly perform.
- In part III, we build on the top of the novel audio-visual architecture and proposed a more complex architecture for audio-visual perception. This architecture is quite unique in its novelty as the following:
  - It is so far the first cognitive architecture that addresses the cross-modal interaction in perception.
  - The biological bases of the architecture as it adopts multiple biological processes and organization in different levels of the architectures.
  - Unlike other audio-visual perception systems, my proposed architecture is generalized to different tasks and not specifically addressing one single task.

  Additionally, we contribute to active sound source localization with a framework for analyzing different head motor strategies to improve the performance of the sound localization system. This framework is actually a sub-part of perception architecture.

  We push further with our architecture and implemented it in multiple robotic platforms and applied the architecture in a real-world application (use case).
- Part IV, is a supplementary part in which I propose two development pipelines for robot audition. The first pipeline is mainly for sound localization and the second one is for environmental sound classification. These pipelines are addressing challenges related to auditory perception in robotics.

2.7.2 Detailed Reviews In The Chapters

For the contributions presented in parts II and III, each of the points is connected to the main aim of the thesis which is pushing toward shared-perception cognitive architecture. In each chapter we will be providing a detailed review of the required background and state-of-the-art methodologies as the following:

- In chapter 3, I will review joint attention and its context in the context of human-robot interaction. While in chapter 4, I will review the working memory role in the attentional tasks as well as the time-variant decision-making process. The reviews in these chapters are the required topics to draw a complete understanding of the proposed architecture that addresses the mutual influence in the human-Robot Joint Attention Task.
- In chapter 5, three concepts will be discussed, dual visual and audio pathways in the brain, perceptual inference, and spatial memory in the context of scene understanding. These concepts are the implemented concepts to address cross-model Interaction. Further in chapter 6, the importance of proactive perception will be addressed, and finally, in chapter 7, the generalization challenge in robotics will be reviewed, and how the proposed perception architecture is applied to different robots and in a real-world application.

Additionally, during my research, I targeted to discover and propose alternative developmental pathways for the robotic audition for the existing common approaches to tackle the challenges in the robotic audition. This will be in chapter 8 and chapter 9. In these chapters, I will review sound source localization in robotics in detail and learning methods for audio classification. Then I will connect the review to the proposed alternative development pathway for audio localization and shallow models and continues learning for audio classification.

In this chapter, I reviewed the relevant points to my research approach. Following this, In the following chapters, deeper reviews will be presented related to relevant topics to each chapter. As the second part of the thesis is focusing on building an architecture for joint attention, in chapter 3, topics related to joint attention will be reviewed. Further, in the same part chapter 4 will review the related work done in the concepts used in the architecture which are the working memory and the time-variant decision-making process.

The third part of the thesis focuses on building architecture for perception, the model is based mainly on three phenomena. What and Where pathways, perceptual multi-model inference, and finally spatial memory and scene understanding. These three concepts will be reviewed in the first chapter of this part (chapter 5).

In the fourth part, chapter 8 will review the related work in sound source localization for robotics.

2.8 Chapter Conclusion

This chapter gives a general overview of the thesis. Firstly, It explains the main motivation of the whole research which is having a human-robot shared perception to promote effective collaboration. It also frames the requirements and research directions toward this aim theoretically. Following this, the chapter gives a brief explanation of cognitive architectures and the reasons for each element in our approach (Biological inspiration, multimodality, generalization, and attention based). After that, the chapter mentions the starting point of the development and the contribution of the research work. Finally, the chapter point to a detailed review of the required background in their corresponding chapters.

Part II

Audio-Visual Cognitive Architecture For Joint Attention

Chapter 3

First Implementation of an Integrated Audio-Visual Architecture For Attention

> Rapport demands joint attention – mutual focus. Our need to make an effort to have such human moments has never been greater, given the ocean of distractions we all navigate daily.
>
> **Daniel Goleman**

Joint attention is a vital skill in human-robot interaction. Modeling the attentional process on robots would help in understanding the attentional process in humans. This chapter introduces an integrated audio-visual attentional architecture based on the PROVISION attention, and the Bayesian auditory localization model. The model follows the latest propositions of integration of visual and audio attention. The novelty of the model is in two aspects. Firstly, it proposes a new method of attention selection that moves to an acyclic technique instead of winner takes all traditional method as found in the latest attention models [13, 176]. Secondly, it moves from the 2D space to the 3D world and connects attention with actions in the 3D world. Further, a preliminary experiment was conducted. The experiment is a joint attentional task between a robot and a human participant to study the attentional timing of the robot (in comparison with the human participant), and the effect of mutual presence using two kinds of stimulus. (Audio only, Audiovisual) The results of the experiment showed good performance when the audio-visual stimulus is presented, the complexity of the task generally and especially for the audio-only stimulation, and the incoordination between perception and action cycle. To address the drawbacks of the performance, the study suggested integrating other cognitive processing and decision-making process.

3.1 Introduction

Joint attention (JA) is defined as the shared attentional focus on the same perceptual event by multiple individuals that coexist in the same environment [158]. The joint attention phenomenon has been studied in cognitive sciences [219] and psychology [63, 101] typically on infants and autistic children [164, 29]. It is also an important social cognitive skill for the development of infants [161, 162, 20]. Additionally, it improves the coordination in the collaborative interactions [27, 206]. This draws the importance of joint attention in human-robot collaborative interactions.

There are multiple levels of Joint attention that are developed in infancy [29, 158]. To understand these levels, a recent typology has been proposed that defined the different levels of joint attention as a level of jointness [228]. The typology defined a scale of four different levels (monitoring attention, common attention, Mutual attention, and shared attention) based on the common knowledge between both agents. Another topology of defining the levels of joint attention is based on the ability of knowledge perception and processing [163] which is associated with the development process of joint attention. This topology defines three levels of joint attention. The lowest level is sharing the gaze with the other agent on the same object which is the attention ability toward the shared stimulus in the environment in both spatial and time perspectives. Followed by dyadic joint attention behavior, which is understanding the social interaction between two agents such as when making eye contact, smiling, cooing, etc. Finally, the highest joint attention behavior is the triatic behavior that involves two agents and the environment. The triatic behavior engages multiple processes with different complexities such as understanding the intentionality of the other agent, initiating a joint attention bid, and maintaining the self and other intentionality. Joint attention is initiated by a bid from one agent such as pointing or directing the gaze towards an object, or from a shared stimulus in the environment with a shared intentional goal for both agents. Yet, understanding the development of Joint attention and its levels is an open challenge [107, 228, 160]. To address the challenges related to joint attention, modeling the joint attention process on robots was strongly suggested in many articles [107, 46].

In recent years, different computational models of attention for artificial agents have been developed [121, 166, 93, 240] to respond to visual stimuli [94], auditory stimuli [105], and audio-visual stimuli [14, 3]. The role of joint attention during the interaction of human and artificial agents has been investigated through the study of attention timing of gaze patterns [185]. On the other hand, other studies focused on the human side of a human-robot interaction scenario and studied joint attention with the use of robotics [37, 118, 251]. However, only a few attention systems have been designed and evaluated to specifically address the context of collaboration between the human and the physically present robot partner [1]. In particular, aspects such as the attention-timing and focus redeployment strategies, are not addressed in the computational modeling of the attention system. This neglect the impact of mutual presence in the human-robot interaction. Such mutual influence has a fundamental importance, especially when assessing joint attention, and yet this dimension is often not considered in joint attention studies. Adding to this, all of the proposed models that address the joint attention were using only visual environment and none of these models considered the multi-sensory environment. This aspect is important as a recent study [16] explained the importance of nonverbal cues on joint attention and how non-verbal cues enhance and extend joint attention capabilities.

Having human-robot joint attention as a motivational objective, the main aim of the work presented in this part of the thesis (chapter 3, and 4) is to build an audio-visual cognitive architecture for attention with the consideration of the existence of a human partner in a joint task. Additionally, to study the mutual influence of such an interaction. In particular, how the action of the robot such as pointing and gaze action would influence the attention of the human partner. Additionally, to examine the attentional spatiotemporal performance of the robot with the existence of a human partner in the scene versus this partner.

The contribution in this chapter has dual objectives: first, to improve the existing multisensory systems for attentional redeployment in order to address joint attention in collaborative tasks, and second, to demonstrate the performance of our solution by evaluating the whole collaborative group comprising both the human and the robot participants.

3.2 Methodology

In our work, we developed starting from the existing PROVISION system: a Proactive saliency-based selective visual attention model which was implemented for iCub robot [201]. In its bottom-up implementation, the attention system decomposes the visual scene into a list of feature maps. The features are linearly combined with a specific weight for each of them. We integrated the PROVISION system with an auditory attention system based on the Bayesian sound localization model developed for the humanoid robotic platform [119]. As well as PROVISION, the auditory attentive system is a biologically inspired model that uses

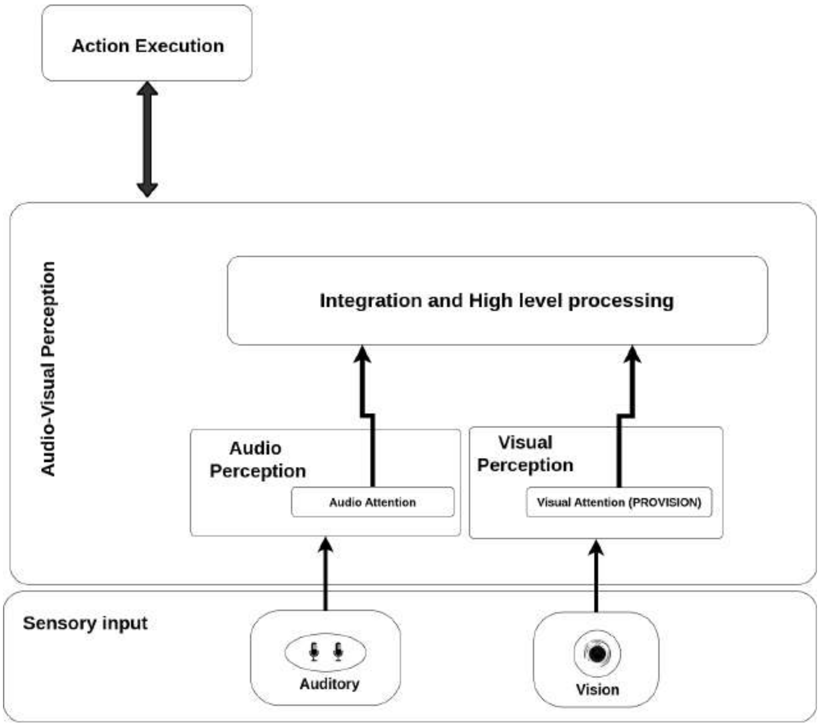


Figure 3.1 The overall structure of the cognitive architecture

only binaural sensing (two microphones in the head of the humanoid robot iCub) in order to calculate a Bayesian allocentric probability map for the location of the sound source. Thus, we extracted an egocentric map from the allocentric output from the auditory localization system and map the angles that are in the field of vision to a cartesian saliency map. The cartesian saliency map is taken as an additional input feature to the spatial linear combination of the topographic saliency maps. The objective is to allow the robot to have an attention mechanism based on the audio source in combination with visual attention and to reinforce the selectivity of the targets by relying on both visual and auditory features instead of visual features only.

Our first new contribution aims at investigating how a biologically plausible multisensory attention system should redesign attentional timing and focus redeployment strategies to promote joint attention between humans and robots. The preliminary step is to move from cyclical selection of attentive focus, typical of attentive systems implemented on robots, to the temporal asynchronous attention selection in sensorial landscapes. This allows for the resembling of asynchronous attentional redeployment of a human partner which in turn facilitates joint attention. This feature is implemented through the acyclic extraction of a saliency “hot point”. A saliency “hot point” is a spatial location in the scene that is extremely salient when compared to the rest of the scene. It is not only the most salient point in the scene, but it also has a high salience compared to the whole scene. This spatial location is what is identified as worth attending by the attention of human interactions when a stimulus is presented. Therefore, we compare the maximum salient point with the whole distribution of attentive responses across the entire combined saliency map. If the value of the maximum point exceeded the triple of the standard in comparison to the mean value with a certain threshold, then it is considered a hot point. It means that this point is extremely salient compared relatively to the full scene. The threshold here represents a tuning parameter for the sensitivity of the system. From this point of the paper, we will refer to (maxValue – meanValue – 3 σ) as " gamma Value".

$$\Gamma = (maxValue - meanValue - 3\sigma)\begin{cases}\text{hot point} & \text{if } \Gamma > threshold \\ \text{not a hot point} & \text{if } \Gamma \leq threshold\end{cases} \tag{3.1}$$

σ : standard deviation of the combined saliency image

The second contribution of this work is the projection of the retinotopic response into allocentric spatial representation for motor control. The motor control system in humans is expecting the allocentric world-based 3D location from the location of the target in the scene to properly respond and initiate motor commands [40]. Similarly, We provide this by using prior knowledge about the contextualization of the environment. It is reasonable to assume that in the context of the interaction is the shared working space. Therefore, we defined the attentive plane as the geometrical plane where the speakers and bulb lie on. Knowing the plane and using the single image the 3D projected location of a point in the given image is expected from the origin of the robot. Mathematically, this is done by computing the 3D projection (x,y,z) of the pixel location (u,v) in the image on the plane defined by the equation "aX + bY + cZ + D + 0". The attentive plane can be represented as the pre-defined region which includes the selected stimuli that require a response in a specific context. It is the area of the cooperative task between the human and the robot. This bit of information is a shared prior knowledge between the robot and the human. Finally, the point will satisfy the response execution if the expected 3D point is in the pre-defined cooperative area of the task. The following equation shows the simple form of the projection 3D image projection where f is the focal length of the camera.

3.2.1 Software Infrastructure

The system is developed using Yarp [150] in C++ and python. In the design of the software infrastructure, we opted for modularity and for multiple connections between modules.

$$\begin{pmatrix} u \\ v \end{pmatrix} = f/z \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{3.2}$$

2D to 3D projection
u: the vertical component of the vector in the image
v: the horizontal component of the vector in the image
z: the pre-defined depth of the plane
x,y,z: the £D location in space
f: focal length of the camera

@default@default

Figure 3.2 Software infrastructure integrating auditory and visual attention systems in order to provide attentive motor commands. The PROVISION model is shown in yellow, the audio Bayesian model is shown in green, and the new components are shown in red.

Figure **??** shows the overall structure of the system implementation and Figure 3.2 showed the detailed structure highlighting the contribution. The PROVISION model is shown in yellow and the Bayesian auditory localization system is in green. Generally, the model is integrating the auditory as a feature in the PROVISION system (Saliency map). The newly added modules/process are shown in red background.

For the audio-visual model, we had to implement an integration algorithm where both visual attention and audio attention are aligned and have the same representation. This integration is designed to be processed in the integration and high-level processing component of the audio-visual perception block. In this component of the architecture, auditory attention is integrated together with the visual attention system. We remapped the allocentric auditory map into a visual egocentric saliency map. The map is then added as a feature to the linear combination of the attention system (already developed in the visual attention PROVISION model [201]). The sound then reinforces the visual saliency map at the corresponding azimuthal location only if the source of sound is located within the field of view. The aim of this process is to provide a unified multi-sensory saliency map that enables the identification of salient points from both auditory and visual signals.

After sensory integration, the output of the integration process is a saliency-integrated map. Next, the saliency selection process happens, in which the system selects the point the model needs to attend to. Therefore, we implemented a temporal asynchronous selection process at salient changes in the landscape of the perceptual sensors (as mentioned in the previous part). This is done in the attention manager component which allows the system to resemble the asynchronous attentional redeployment of humans.

These added modules are explained as follows:

Egocentric audio cropper in this module, we generate the cartesian saliency map from the allocentric audio Bayesian map. The allocentric audio Bayesian map can be represented as an array of 360 probabilistic values corresponding to the 360 degrees centralized across the axis of the robot. Based on the current head azimuth angle and the gaze angle we cut only the relative angles that represent the field of vision (FOV). If the maximum probability of the sound source location is among the field of vision exceeded a certain threshold a strap is created in that location and then extended to a 2D cartesian image. This Cartesian image is then sent as a cartesian feature map to the PROVISION attention model. In this architecture, we present the maximum probability as the confidence value of the cartesian image of the audio. o achieve this task, the module needs to know the current state of the locations of both the head and camera in the azimuth direction. The process of extracting the egocentric map is based on the current locations on the camera and head in the azimuthal direction and the camera parameters. The camera parameters specify the width of the area of vision, while the location states of the camera and the head specify the middle value in the area of vision range. Knowing the middle angler value and the angular width of the sound source, the module computes the starting and ending degree angles which then are extracted from the allocentric map. This is the first stage of the audio egocentric module which has an output of a subset from the allocentric saliency map of the audio. The second stage involves scaling these values vertically and horizontally to be in equal size with the frame size of the visual image. The horizontal scaling assumes that the audio source is from the horizontal level in the scene as we only consider the azimuth plane in the audio localization module. The output of the scaling stage is now ready to be integrated as a feature in the PROVISION attention system with a defined weight in the linear combination part.

Attention Manager This module is a controller for the whole attention system. It is also a middle module between attention and any other applications or modules. It is responsible for the computational process of the hot point as well as sending/ receiving commands with external systems. It receives the combined cartesian saliency image after the linear combination and also communicates with the action execution part with the required information. This module has full control to suspend and resume the attention process as well as manipulate the parameters of the PROVISION system as well as the audio integration stage. This module represents the gate for any external modules that need to communicate with the attention system.

Attention Action Linker is the module that estimates the 3D location of the point of interest in the scene and checks if it is inside or outside the attentive plane. It gets the hot point from the attention manager module. It is also communicating with the attention manager and sharing the state of the action. This module limits the action execution to be done under certain pre-defined conditions. The conditions are location constraints in the 3D world. We build this module as a decision-making layer to decide when and how the action will be executed. Additionally, it sends the state of the action execution to the attention manager so that the manager is able to suspend the attention system and the gaze during the action execution and resume them after finishing the action.

3.2.2 The Experiment

The aim of the experiment is to examine the performance of the robot's attention using the proposed architecture on a joint task with a human participant. Thus, the existence and the actions (movements) of the human participant affect the attentional system. Also, this setup will allow the study of the robot's effect on the participant's attention. This experiment is the pilot study which is mainly aiming to understand the dynamics of the proposed architecture and further develop the required improvements to study the mutual effect between the robot and the participant. The analysis is done by applying two kinds of sensorial stimulations. The first one is based exclusively on auditory stimuli only and the second one combines audio-visual stimuli. We used the humanoid robot iCub [151] in our study. It is equipped with different sensors including two cameras (eyes) and two microphones (ears) which we used in our experiment to perceive the environment. In the experiment, the subject is sitting on a chair facing the robot. In between the subject and the robot, there is a black table with the stimuli distribution on it. The audio stimuli are produced using four identical black Bluetooth speakers distributed in a horizontal line fifteen centimeters apart. The audio stimuli are 240 Hz sin wave. The distance between the speakers' line and the robot is 65 centimeters while for the human is 45 centimeters. The first speaker on the left side of the subject is coupled with a colored bulb which represents the visual stimuli. In this experiment, we used a fixed color (blue). Figure 3.3 shows the setup of the experiment.

The experiment consists of 32 rounds for each subject. In each round, a random number is selected between one and four corresponding to the location of the stimuli. Then the stimuli are activated exclusively for the selected speaker. If the round is running for the first speaker then the bulb will turn on in a synchronic way with the speaker. The turn-on duration is 10 seconds. The subject is requested to react as fast as possible by pressing the two buttons

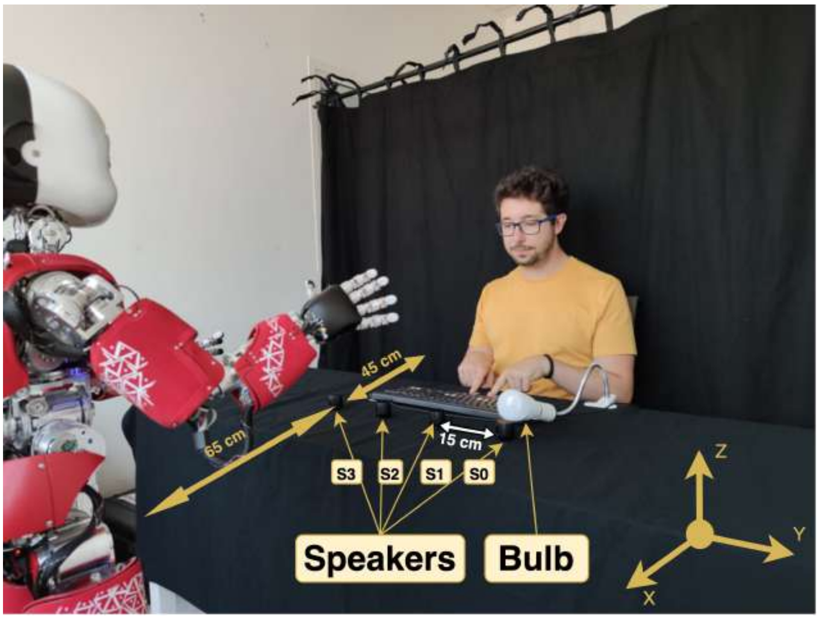


Figure 3.3 Experimental setup

corresponding to the activated stimuli using both of his/her index fingers and returning them back to the initial position on the edge of the table. The buttons in the keyboard are selected to be approximately equal in distance from the initial position. The time between the round is 10 seconds so the total time of the round is 20 seconds. The location pattern is randomized but with the same sequence for all subjects as the following sequence: (S1, S0, S1, S3, S2, S2, S0, S3, S1, S3, S2, S0, S3, S2, S3, S1, S0, S2, S3, S0, S1, S0, S0, S2, S1, S3, S1, S2, S0, S1, S3, S0).

3.2.3 Measurements

The main goal is to compare the attention performance of the human and the artificial agents in the same cooperative task. Therefore, and for the human side, we recorded the reaction of the human in which keys are pressed. For the robot, we recorded the profile of the audio as well as the full attention system and the action execution commands. The profile of the audio consists of the value of the maximum confidence as well as its egocentric location. From the full system, we recorded the analysis of the combined scene to have the gamma value and its expected location in the 3D world.

3.3 Results and Discussion

In order to compare the performance of the robot and related with the performance of the human in the joint attention task, we analyzed the overall performance of their temporal response after the stimulus is presented. The stimulus presented is characterized as auditory only when the auditory stimulus is reproduced by one of the speakers and characterized as audiovisual when the stimulus is produced by the light of a bulb and the sound from the speaker. We considered the time window of 20sec representing the cyclic stimulus production. In the cycle, the first 10sec the stimulus is provided in the first 10sec whereas, in the second 10sec, no other stimulus is produced in the scene.

3.3.1 Overall performance

The overall performance of the humanoid robot is compared with the response of human participants. This result is shown in figure3.4 (shows the error count), and figure 3.5 (shows the creation time). In particular, we indicate the number N wrong as the number of false attended locations L attended computed according to the following formula: N wrong = N wrong + 1, if |L target - L attended | > θ. The threshold θ = 0.10[m] since the distance between two consecutive stimulus locations is 0.15[m]. For the robot, it is computed knowing the fixation point commanded by the attention system and for the human participant, it counts the number of wrong stimulus selections at the keyboard. The number of wrong-attended locations is comparable only for trials where a visual-auditory stimulus is presented whereas there is a significative difference between the performance of the robot and the human participant when the stimulus is exclusively auditory. It is worth noting that especially for auditory stimulation the task is difficult for the human participants and the robot since the auditory localization of pure tone without head movement provided within an azimuthal angle range is perceptually challenging.

Comparing instead the performance of the robot and the human participants on average, it is worth noting how the robot has wider variability in response time RT robot but the response time is not significantly different with respect to human participant‘s response time RT human. This is also true if we discard RT human > 5[s] as human subjects’ mistakes. In fact, if both keys failed to be synchronously pressed the system does not record the human‘s response. The robot instead shows an immediate response indicating quick detection of the presence of a new stimulus but it is also not as accurate as a human participant in localizing the stimulus source in space.

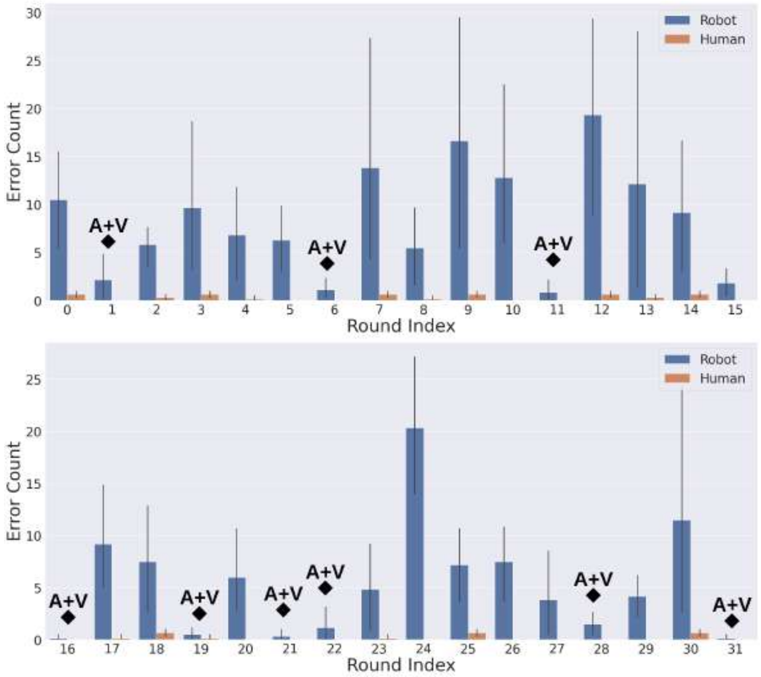


Figure 3.4 The Error Count for both the robot and the human participants. The marked trials with (A+V) are the audio-visual stimulation trials which were executed with S0 using both the bulb and the speaker)

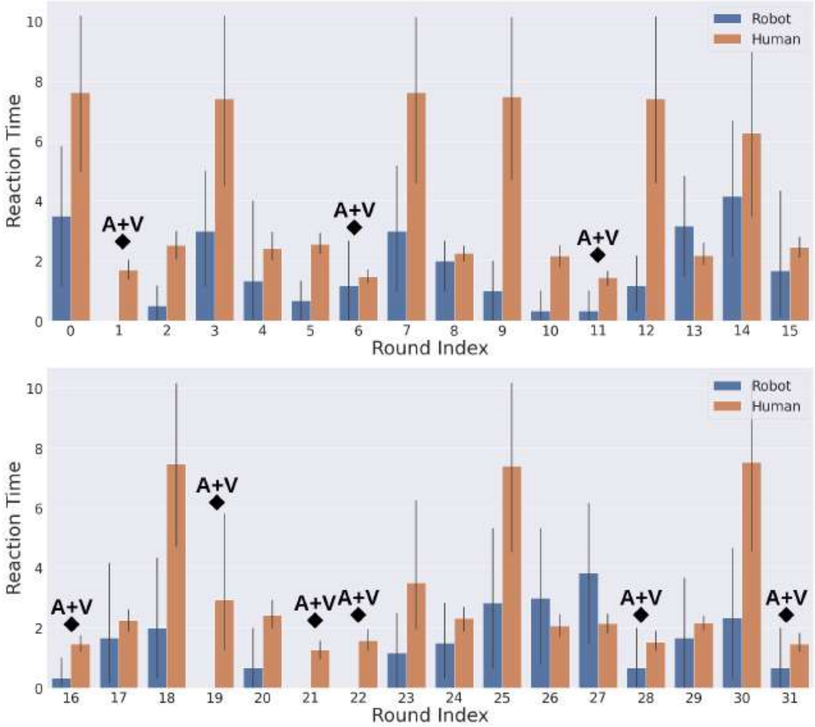


Figure 3.5 Reaction time in seconds for both the robot and the human participants

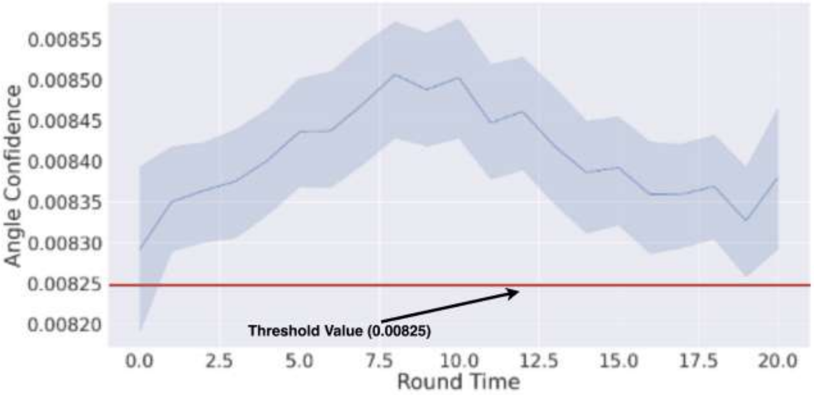


Figure 3.6 Confidence value γ profile. In the figure, an increase during the first 10 seconds (when the stimulus is on), and a decrease after that (when the stimulus is off)

3.3.2 Temporal analysis

For the temporal analysis of the auditory attention behavior of the humanoid robot iCub the value γ indicates the confidence that a salient auditory stimulus is present in the scene. In average figure 3.6 shows how the value γ changes during the 20 seconds of the trial cycle and in particular how the confidence increases for the first 10 seconds and decreases for the last 10 secs of the trial. All the values of γ that exceed the threshold TH indicate the presence of the salient auditory stimulus and the localization process of the auditory stimulus is initiated.

The value of TH is a fundamental parameter that impacts the number of errors since a lower TH will activate localization even when the Bayesian network is not confident enough about the stimulus location but also impact the reaction time since a higher level will prevent the system from attending the auditory stimulus. In the correct implementation, we opted for TH = 0.0082 which has an impact on fast reaction time but also on the high number of errors. For the temporal analysis of the visual attentive behavior of the humanoid robot iCub involved in an attentional task, we focus on the 20 seconds after the stimulus on-set observing in particular how the attentive system responds in presence of the stimulation (first 10 seconds) and in absence of the stimulation (second 10 seconds).

In figure3.7 we report the average of the Γ progression in the 20 sec time window for all the subjects and all the trials. In particular, the Γ value, which is the value that indicates the uniqueness of the most salient stimulus, increases for the first 10 seconds and decreases in absence of the salient stimulus. When the Γ value exceeds the threshold, the multimodal attention system detects the presence of a unique salient stimulus significantly different from the rest of the scene and it proceeds for the 3D spatial localization of the target stimulus. It is worth observing the swift increment of confidence in the average Γ value (<500ms on

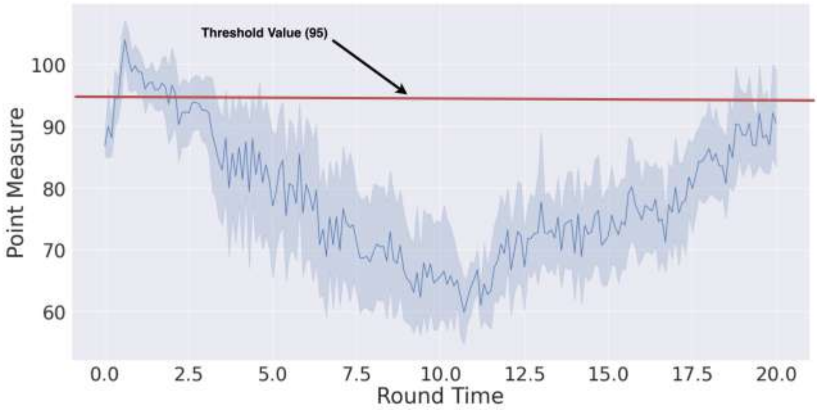


Figure 3.7 average gamma value Γ (±st.error) temporal profile for all the trials involving visual-auditory stimulation. There are three observations. Firstly, a swift increase in the first second (exceeding the threshold). Secondly, an increase after the 10th second. Finally, exceeding the threshold for some trials around the 18th second.

average) that explains the faster response rate RT robot with respect to the RT human. In absence of the stimulus (after 10 secs) the Γ increases because other stimuli in the scene become salient with respect to the rest of the scene. The process is gradual but the Γ value exceeds the TH Γ on average after 18 secs generating new responses of saliency and consequently new target localization attempts. This is where most of the wrong attended locations are generated thus explaining the greater number of errors in stimulus localization.

Concerning the process that spatially determines the position of the most salient stimulus in the scene, the performance is especially promising for the localization of visual-auditory stimuli. Figure 3.8 shows the error along the y-axis (the direction of the speaker deployment and azimuthal angle for the robot) since the x (depth) and z (elevation) axis do not show significative changes. As shown in figure 3.8, as soon as the stimulus is set the error Error y drops to 0.05[m] in less than 500ms. Such (RT robot) observed even before is comparable to reaction time with respect to the (RT human). On average for the first 5 seconds, the error oscillates around 0.05[m] and only partially adjusts in the second 5 seconds remaining on average below 0.10[m]. After the stimulus stops (after 10secs) the robot attends other locations not necessarily corresponding to the target of this trial. Such behavior of the robot in absence of salient stimulus is comparable with the unconstrained behavior of the human participants who in absence of stimulus attends other salient locations in the scene (not necessarily the target of the trial) and sometimes even the robot partner.

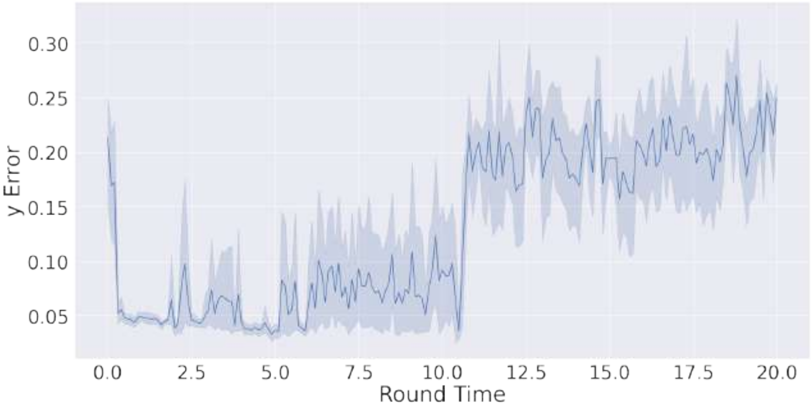


Figure 3.8 Localization error for an attentive task involving visual-auditory stimulation. A decline in the localization error can be observed in the first 10 seconds (when the stimulus is on)

3.3.3 Discussion and Future Work

The architecture proposed for audio-visual attention in the context of human-robot joint attention tasks shows promising behavior although not comparable with the attentive pattern of the human counterpart. It is clear that, whereas the visual attention system swiftly and correctly captures changes in the visual field that correctly drive the attentive system, auditory attention remains a challenge for both the human participant and the robot. In particular, the majority of the subject mentioned the difficulty in localizing the source of sound when only the pure tone is produced. The interesting question for our future study focuses on the test with complex tone sounds or speech which is richer in terms of auditory features for sound recognition and localization. Whereas the human participant can refine the estimation of the stimulus location by head rotations and then solve the auditory problem [30] the robot does not fully leverage its motor capabilities to disambiguate the uncertain situation. Therefore, we plan to endow the humanoid robot iCub with motor control finalized for the refinement of its estimation and to ask the human participants to wear an eye-tracking system that can give insights into the degree of involvement of head rotation and eye fixation process. Further, the auditory attentive system can be automatically adapted to the contextualization of the specific task. It is more efficient to reallocate the limited computational resources of the robot in the direction of assessing exclusively what is happening in the task (e.g.: limiting the auditory beams to the exclusive area in front of the robot, adjusting the frequency bands to the range of interest) and this can make the robot more efficiently react to the auditory stimulation. In fact, the task for the human participant is to locate the stimulus from one of four known locations but this is not the case for the robot as it tries to solve the localization problem across all azimuth angles. Another interesting point is that multisensory information equally supports the attentive process of both the human observer and the robot observer. In situations where both the auditory and visual stimuli mutually reinforce, both the human partner and the humanoid robot iCub improve their attending performance. This indicates that a richer testbed with multisensory stimulation (auditory and visual) for all the stimulus locations in our following experiment might give us insights into the benefit of multisensory integration across different locations. The last point to discuss is the robot's system stability. From the results and the observed behavior during the experiment, it is clear that the robot's reaction is not stable. In particular, the robot reacts multiple times to the same stimulus and also reacts to other stimuli in the scene but unrelated to the task. For example, the participant's movements. This is due to the fact that the robot doesn't track the stimulus and also doesn't track his own action. This suggests that other cognitive components might be required to endow the robot with the capability of performing this task in a stable way.

3.4 Conclusion

If robots are going to be used to support daily activities, it is important to understand how the process of joint attention work in typical human-robot interactions. Joint attention is an important mediator for efficient collaborations between the interactants however, it is challenging to endow robotic platforms with such complex cognitive capabilities. This is also due to the challenge of sharing the same attentional timing and precision capabilities (reaction time and localization accuracy) as human partners. Our contribution aims at improving existing auditory and visual attention systems with specific mechanisms that promote attention in human-robot collaborations. We focus specifically on the examination of the attentional mechanism in a joint task. We demonstrated that the performance of the proposed system is comparable with the performance of human participants when the multisensory stimulation (auditory and visual) is presented at the same time to both the human participant and the robot. On the other hand, the challenge of attending only auditory stimuli is only partially fulfilled for the humanoid robot since the robot correctly recognizes the presence of salient new stimulation but shows limitations in the correct localization of the stimulus. This limitation in the auditory channel is generally due to the complexity of the signal and the time the bayesian system needs to converge to an acceptable level of confidence. Another important aspect that is important to mention, the system doesn't differentiate between new stimuli and already detected stimuli. Indeed, the architecture wasn't designed to track the stimulus but this is an important point to close the loop between perception (attention in this case) and action. Therefore, it is suggested to integrate other cognitive components such as memory and decision-making components to track the stimulus and also to address the tradeoff between accuracy and time which is generally in auditory processing due to the required time of reaching an acceptable level of confidence. In this chapter, we proposed an audio-visual cognitive architecture and evaluated the performance of this architecture on the iCub humanoid robot in a joint task with a human participant. The experiment showed promising results but also some limitations. In the following chapter, we present how we addressed these limitations.

Chapter 4

The Extended Version of The Integrated Audio-Visual Architecture Using Memory Based Decision Making

The true art of memory is the art of attention

Samuel Johnson

The results of the experiment presented in the previous Chapter (3) suggested the need for other cognitive components to coordinate the actions that are executed toward a represented stimulus in the scene. Also suggested the need for a decision-making process that addresses the time-accuracy tradeoff. In this chapter, an expanded version of the architecture is proposed taking into consideration the propositions from the previous experiment. Taking a biologically inspired approach, the architecture added two main biological components to the previous architecture. The first component is time-variant decision-making, and the second component is working memory. Further, an extended version of the previous experiment was conducted using eye-tracking glasses to track the human gaze during the experiment and a new set of identical audio-visual stimulation devices. The experiment studied the robotic and the participant's attentional behavior, and how the robotic action (gaze + pointing) is affecting the attention of the human participant. The results showed that the new version of the architecture stabilized the attention and action cycle and was capable of tracking the stimulus and the robot was capable of tracking its own behavior. For the attentional timing, the robot's attention was comparable to the human's performance in the audio-visual stimulation condition. However, it was relatively slower in the auditory-only condition.

4.1 Background

4.1.1 Attention and Working memory

Working memory has been defined as short-term memory used in order to reinterpret information to operate better in the environment proactively. It temporarily stores and manipulates important information for complex cognitive tasks. The term was first introduced by Miller, Galanter, and Pribram in 1960 [154] in cognitive psychological studies and further adopted and used in other research domains including neuroscience and cognitive robotics. Researchers from psychology, neuroscience, and cognitive sciences widely agreed and showed the interaction between the attention process and working memory. They demonstrate the influence and the interaction between the working memory and the attention mechanism in different ways [11]. For selective attention, working memory is argued to contribute to the control of attention by holding the relevant information of the perceptual task, controlled action, and templates of the targets [175]. However, this interaction is rarely addressed in cognitive architectures, especially in the collaborative robotics field. The different models proposed for visual attention and auditory attention for robots didn't consider the working memory role in the attention focus redeployment. We conclude that the role of the working memory in the attention cognitive architectures for robotic agents isn't yet addressed.

4.1.2 Time Variant Decision Making Process

From research theories elaborated in the previous decade, visual processing in humans and animals triggers a decision-making mechanism in the form of a higher-level process, relying on the extraction of low-level features and properties from visual input [244]. This process is meant to evaluate the perceptual output properties and their relevance to the current goal and expectations.

Decision-making processes inspired by time-invariant models have been adopted for decades by the computational neuroscience community[199]. These models are based on a decision-making signal, which is triggered by a fixed threshold. The process integrates confidence over time and once the confidence reaches the fixed threshold, the decision is made and the signal is executed. Recent studies [165] [49] [39] [215] have shown that the time dependency of the decision-making process and the urgency of signals are invoked by humans. These findings show that humans may make decisions with different levels of confidence based on urgency. The more urgent the decision, the less confidence may be accepted. This urgency-based process allows humans to adopt time-variant pressure to execute actions (execution pressure) as a time-variant variable. The first study also showed the existence of neural gain modulation for urgency generation in humans, which implies the existence of modulation signals. These signals are initiated to express urgency and modulate the confidence level.

4.1.3 Aims and Approach

In this work we intend to endow the robot with the ability to rely on working memory, to reinterpret the information acquired in previous instances and states in order to better attend to the environment. Different possible computational models of working memory have been provided in different cognitive studies [204] and in robotics applications [186]. Inspired by these previous works, we provided the robot with a simple implementation of working memory that improves the attentive performance of the cognitive architecture for the humanoid robot iCub [152]. The implementation engages the working memory component in a biologically inspired decision-making process.

Thus, we propose and evaluate the performance of a computational cognitive architecture for memory-based multi-sensory joint attention. Our goal with this study is to validate emergent joint attention guided by our cognitive framework. The architecture includes a multi-sensory attentional model, a working memory, a decision-making element, and an action executor (motor controller) to solve audio-visual stimuli localization with human-like performance. We implemented a bio-inspired decision-making strategy [165] for multi-sensory integration that will take into consideration both cognitive models of attention and the processing of working memory. We aimed at studying how the cognitive architecture responds in collaborative tasks between the iCub robot [152] and a human partner. We address the concept of joint attention emerging from a biologically-inspired multi-sensory selective attentional process defined as the selection of the relevant stimulus while ignoring irrelevant stimuli in the current environmental state [174]. With the goal of endowing an artificial agent with the ability to attend to salient objects as humans do (accurate in location estimation and with optimal timing), we can promote emergent memory-based joint attention in collaborative scenarios. To evaluate the attentional performance in a joint task during an unconstrained interaction (where the action of the human might affect the attentional system) and to exploit mutual influence between the robot and the human participant, we compared human performance with the robot performance in a task in which both agents are exposed to the same salient audio or audio-visual stimuli. In particular, we focused on decision-making as our main contribution, and we then addressed perceptual performance (localization accuracy and reaction time) during the task. Our main testing and performance analysis is structured around three main hypotheses:

H1: **Memory-Based Decision-Making Process:** Empowering the cognitive architecture with a memory-based decision-making process will stabilize the attentional performance.

H2: **Audio-Visual vs Audio Only:** The stimulus localization accuracy and reaction time of the robot in audiovisual tasks is better than in audio-only tasks;

H3: **Robot Performance** The performance (accuracy and reaction time) of the robot will be as good as the performance of the human participants in localizing the stimulus.

4.2 Methodology: The Proposed Architecture and Its Implementation

Getting inspired by the two biological processings which are the Working memory role in attention and the time-variant decision-making process, we proposed a cognitive architecture for audio-visual attention shown in Figure 4.1. We designed this cognitive architecture with three main contribution goals to the scientific community. The first goal was to build a multi-modal (audio-visual) attention computational system to facilitate joint attention between a robot and a human during an interactive task. The second goal was to address the accuracy-time trade-off in attentional decision-making inspired by human behavior. The third goal was to improve the attention, decision-making, and action execution cycle by including a working memory component. The first goal relates to the audio-visual perception component while the second goal concerns the decision-making process. Finally, the third one addresses the role of working memory in the decision-making process and the stability of the attentional system. We also contribute to the research community with the implementation of this novel architecture and applying it to operate in real-time on robots.

Our Cognitive architecture is composed of four main building blocks. In this section, we will explain in detail the four blocks (Audio-Visual Perception, Decision Making, Working Memory, and Action Execution). The details will include the biological inspiration, the overall process, and the connections between the different blocks. The perception block uses early features from both the sensory inputs (the audio and the vision) to trigger the start of the decision-making process. The decision-making process modulates perception to meet the task requirements and further sends commands to the motor control for action

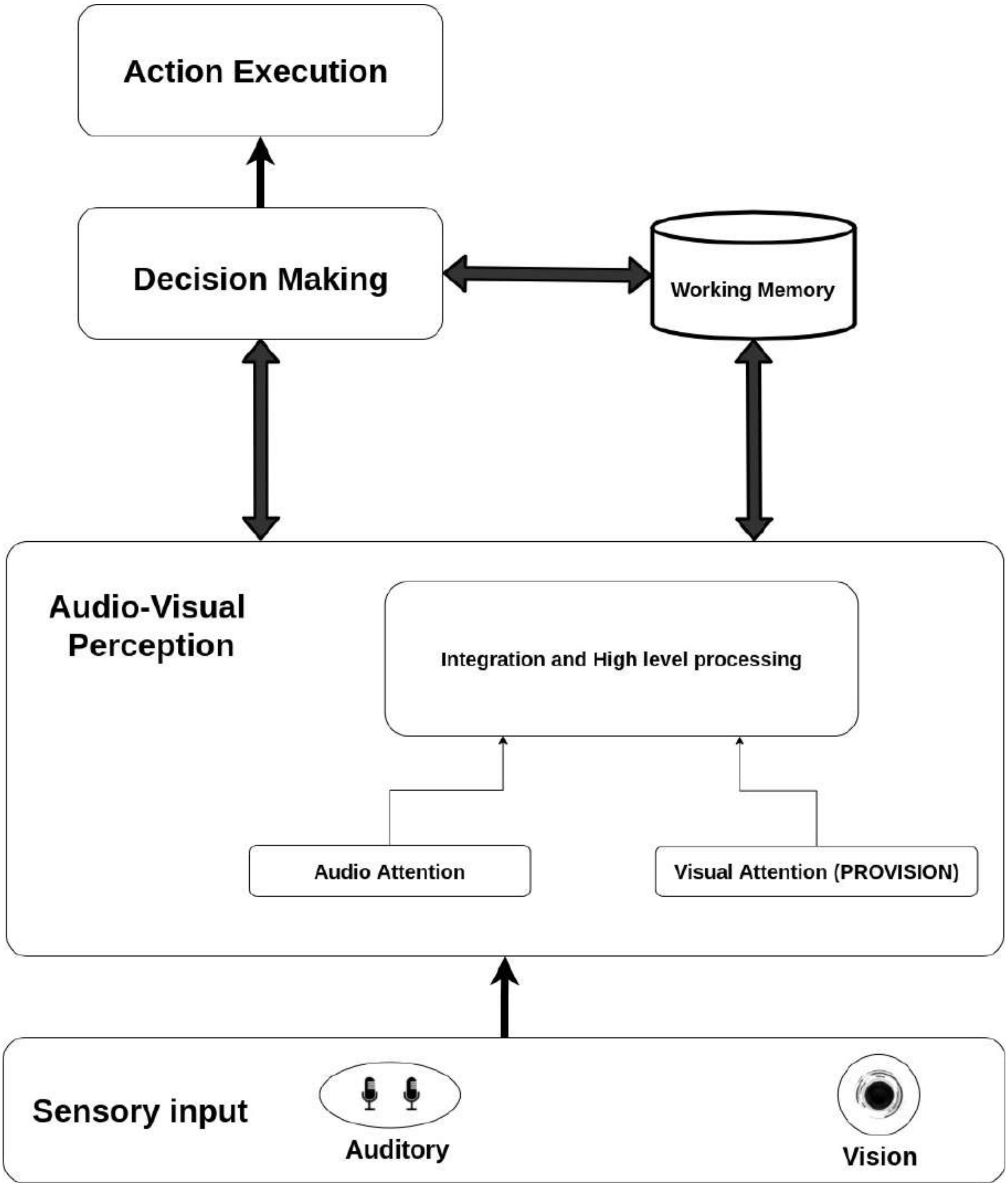


Figure 4.1 Overall cognitive architecture with all the main layers

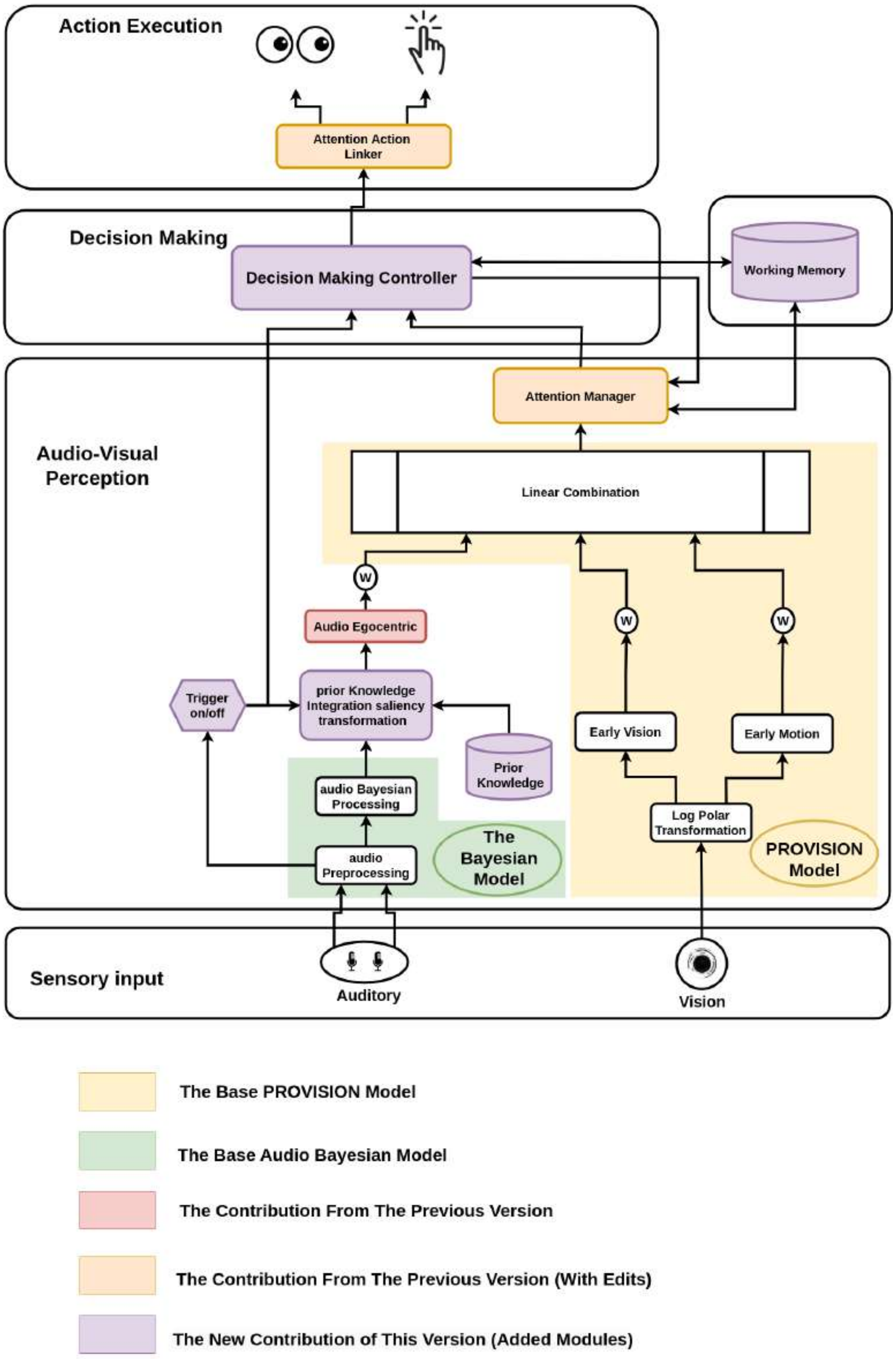


Figure 4.2 The detailed representation of the proposed architecture. The colors of the modules indicated the contribution of this version in blue, the edited modules from the previous version presented in chapter 3 in orange, and the contribution of the previous version (with no edits) in red.

execution. Finally, the memory governs the entire process and is shared between all of the units. We will also explain the technical implementation for each component of the cognitive architecture after mentioning the overall functionalities of the component. Figure 4.2 outlines the structure and connections of our model's modules.

4.2.1 Audio-Visual Perception

The integration of audio and visual modalities is done exactly the same as in the previous version presented in chapter 3. The output of the audio Bayesian localization model is transferred to an egocentric saliency map which is then added as a feature in the PROVISION linear combination stage. The new components of this part are the power trigger and the prior knowledge integration. However, this integration is only performed when the sound exists (the power of the sound exceeds a certain threshold)

The second difference in comparison with the previous version is the attentional selection mechanism. Instead of a fixed threshold Th for the gamma value Γ, the threshold is time-variant in this version. This threshold is defined based on a confidence-urgency trade-off from the decision-making block (a new block in this version which will be discussed in the following section). The Audio-Visual Perception block is also connected to the working memory, in order to update the perceptual states in the memory for a better memory-based decision-making process. In this process, a confidence-urgency trade-off is performed based on the time state and the stimulation states. More details concerning this will be discussed in 4.2.3 Working Memory.

Another added component in audio-visual perception is the integration of prior knowledge for audio perception. The prior knowledge is the spatial locations of possible stimulation sources. This knowledge influences the perceptual abilities of the robot. This process is inspired by biological evidence about the importance of prior knowledge in decreasing cognitive load, improving learning abilities, and improving perception [45, 43]. In Figure 4.2, the PROVISION model is highlighted with a yellow background color, and the audio Bayesian model is highlighted with a green background color. The following part of this section is explaining in detail the implementation of the added components to the audio-visual perception block which was mentioned above in brief.

Trigger, and Prior Knowledge Integration

In order to overcome false positives coming from ambient sound in the environment, we integrated a power detection algorithm along with our sound localization system as a relevant

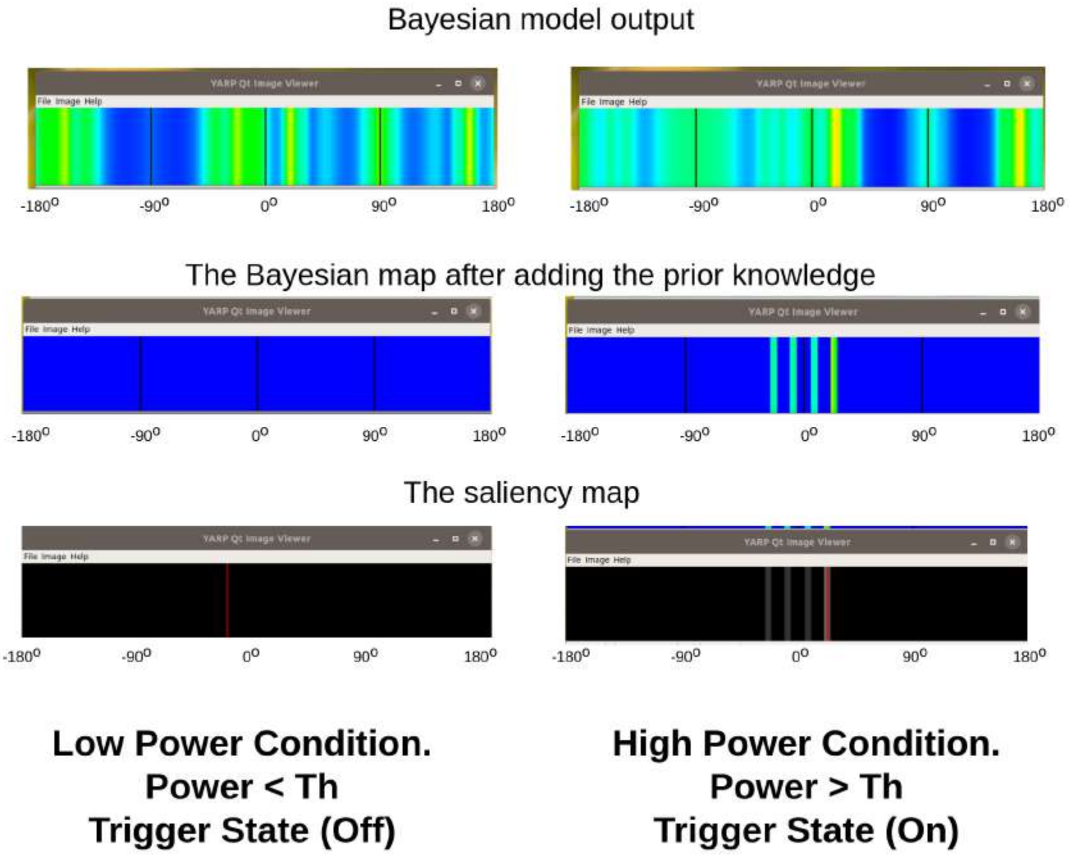


Figure 4.3 The visualization of the trigger and the prior knowledge integration for the auditory maps. The right side figures show the process when the trigger is on and the left side figures are when the trigger is off. The Bayesian map is colored from blue (probability = 0) to red (probability = 1).

attentive mechanism in human audition [214]. We aimed to test the reliability of the sound power as an early informative feature. We added the calculations of the sound power in an early stage (audio prepossessing module) of the audio input. Using a fixed threshold on the total power for both audio channels, the system can determine whether the audio signal is high enough to be considered a valid sound or is just ambient noise. Indeed, a relative threshold would be more efficient in handling the noisy environment but for simplicity, we used a fixed threshold that is autonomously extracted from the environment in the calibration phase (assuming a fixed power level for the noise which is a true assumption for the experimental phase). However, an adaptive power trigger is adopted in a later part of the thesis (chapter 7). The instantaneous sound power is used as an input to the trigger block. The trigger module receives the audio power processed by the audio prepossessing module. Based on a defined threshold for the instantaneous power, the trigger outputs signal to a higher level audio perception module (Prior Knowledge integration & saliency transformation) and also to the decision-making block. Additionally, it updates the working memory which will be explained in a separate section.

Moving to the prior knowledge integration and saliency transformation module, we define two aspects of prior information for audio stimulation. The first aspect is the possible locations of the stimulation. As the current audio system only considers the azimuth angle, this information is in a form of two lists. The first list is of angles describing where in azimuthal space the audio stimulation might be occurring and the second list is the spatial resolution of the angles, which reflects the size of the stimulation source. Thus for each stimulation source in the scene, we express the location in azimuthal allocentric angles from the robot's head axis as (X degrees +/- resolution). These angles and their resolutions are the only locations that are considered from the allocentric probability map and the rest are ignored. The allocentric probability map is the output of the audio localization model, which is a set of 360 values that represent the probability of the sound source's location at any arrival angle around the robot. These probabilistic values correspond to the 360 degrees centralized around the head axis. After considering the prior defined locations only, the resulting map is normalized to keep the Bayesian representation in the form of a probability distribution. By integrating this prior knowledge, we force the model to only focus on pre-biased defined locations. The prior knowledge is generally added in a top-down fashion from a higher level of perception which can be from vision or other modalities. In our implementation, we give architecture the prior knowledge as a parameter (a list of allocentric azimuth angles that present the possible location of sound sources in the scene). This is valid in our case as the experiment we conduct doesn't involve changes in the robot's position in the environment. The second prior for the audio stimulation is the stimulation audio power. It is used to identify the threshold level of the sensitivity of the trigger. The trigger gives a high output if the audio power exceeded the threshold, which is the defined stimulation power level. Conversely, the trigger gives a low output if the audio power is less than this threshold. This signal is used to activate the transmission of the Bayesian map after adding the priors to the next stages. Otherwise, the transmitted map is a zero map. The trigger supports the prior knowledge module with the trigger signal to activate and deactivate the map transmission.

The next process is saliency transformation. The input of this process is the resultant Bayesian map after adding both priors (the stimulation activation level and the sources angles). The whole map is then multiplied by the total audio power and a scale factor. The audio power multiplication gives more importance to high stimulation than low stimulation (both are above the threshold level) and the scale factor transforms from Bayesian values

(0-1) to the values of the monocular image (0-255). Figure 4.3 visualizes the process of integrating the prior knowledge and the transformation to saliency maps.

Attention Manager

The attention manager is a central control module. We used an enhanced version of the previous version. It receives a combined scene (the saliency map output of the linear combination module of the PROVISION). It is also responsible for analyzing this combined scene. The analysis is basically computing a confidence level (the gamma value Γ). We use the previously proposed novel approach (in the previous chapter) to recognize the unique target point of the scene based on gamma value (Γ)measurement. The Γ value gives information about the confidence level of uniqueness. Higher values are more likely to be a unique target whereas low values mean that in the scene there are multiple salient points with similar levels of saliency. When a unique target is recognized ((Γ) value is greater than the current confidence threshold), it sends the selected point to the next connected elements in the architecture which is the decision-making controller in the decision-making block.

Additionally, the attention manager block receives manipulation commands for the threshold value from the decision-making layer. The threshold here represents the level of confidence in which action is required. Therefore, the attention manager here can be presented as a trigger that acquires an action execution process for that current scene from the decision-making block. Also, the module is able to fully control the process of suspending and resuming the attention process as well as the linear combination parameters. To summarize this part, the attention manager presents the main control unit of attention. It has the ability to change the attention parameters (weights of the linear combination and the threshold). It receives commands from the decision-making component, communicates with the working memory and the PROVISION model, and updated the working memory with combined information about the current scene.

4.2.2 Decision Making

From research theories elaborated in the previous decade, visual processing in humans and animals triggers a decision-making mechanism in the form of a higher-level process, relying on the extraction of low-level features and properties from visual input [244]. This process is meant to evaluate the perceptual output properties and their relevance to the current goal and expectations.

Decision-making processes inspired by time-invariant models have been adopted for decades by the computational neuroscience community[199]. These models are based on a decision-making signal, which is triggered by a fixed threshold. The process integrates confidence over time and once the confidence reaches the fixed threshold, the decision is made and the signal is executed. Recent studies [165] [49] [39] [215] have shown that the time dependency of the decision-making process and the urgency of signals are invoked by humans. These findings show that humans may make decisions with different levels of confidence based on urgency. The more urgent the decision, the less confidence may be accepted. This urgency-based process allows humans to adopt time-variant pressure to execute actions (execution pressure) as a time-variant variable. The first study also showed the existence of neural gain modulation for urgency generation in humans, which implies the existence of a modulation signal. These signals are initiated to express urgency and modulate the confidence level.

Inspired by the biological evidence of the time-variant decision-making processes, we propose a model for the decision-making process that recruits a time-variant decision-making signal with the aid of working memory. The model performs four main tasks as the following:

- The first one is tracking the changes in the working memory and listening to the trigger to detect the state change of the stimulation.
- The second task is threshold manipulation based on urgency. This second process is the main element that addresses the time-variant feature of the decision-making block.
- The third task is analyzing the relevance of the received status within a predefined context and defining the urgency. At this point, the context has to be given for the model as input. In our implementation, the context is the defined spatial working area. The stimulus is urgent if it lies within this defined working area.
- The last task is sending the action execution signal to the action execution block based on the required actions which are also defined in the context.

These tasks are defined in three parallel processes. The first process is reading the states from the working memory, the second process is doing the threshold manipulation threshold, and the third process is checking the context, updating the threshold manipulation with the required urgency signal, and sending the action signals. These parallel processes are implemented in the decision-making controller module which will be described in detail in the following part.

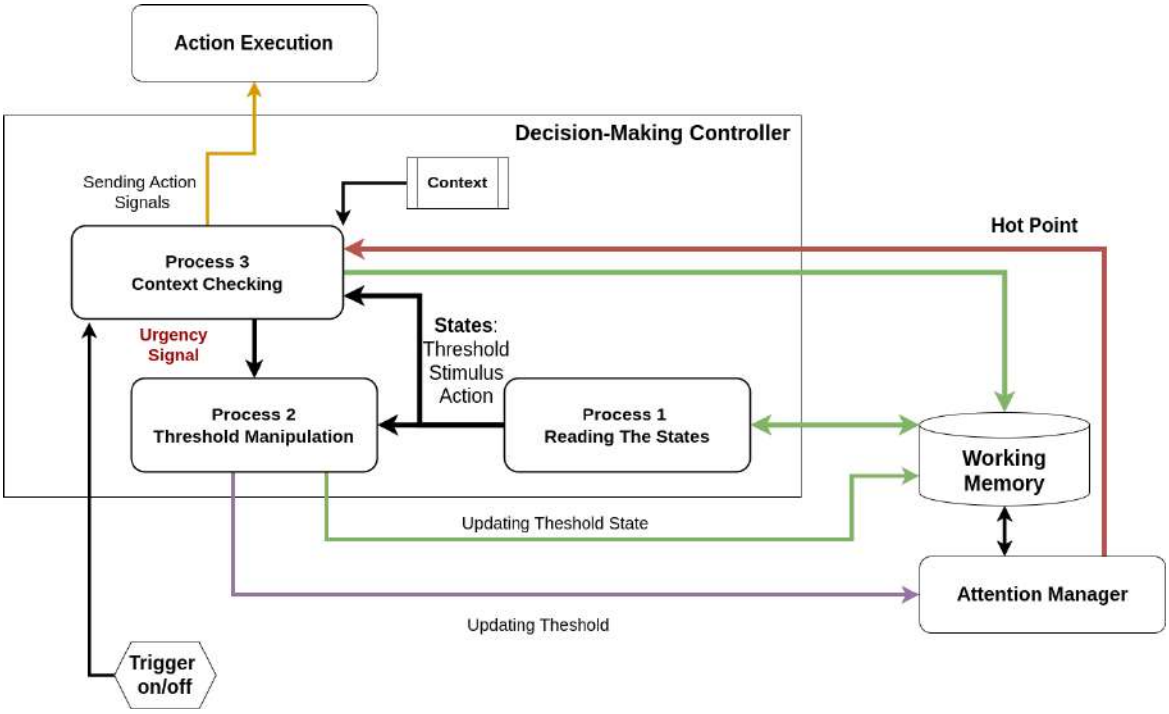


Figure 4.4 The Detailed Process of The Decision-Making Controller Module showing the connections with Attention Manager module, Working Memory, and Action Execution Block

Decision Making controller

The decision-making controller block is the module responsible to control the flow of decisions, manipulating the threshold of the confidence level in the attention manager, analyzing the salient perception output based on the context, and finally sending the request to the action execution system. Figure 4.4 is showing the detailed implementation. The control flow consists of three parallel processes explained as the following:

Process 1: Reading The States This process is periodically reading from the working memory. It reads the states of the stimulus including gamma value Γ and the stimulus position in the image plane. It also reads the current threshold level as well as the state of the action execution.

Process 2: Threshold Manipulation This is the main process of the module. In fact, this is the main novelty of the architecture. It reads the current state of the threshold which is streamed from the first process (Reading The States). Based on an urgency signal, which is perceived from the third process (Context checking) the threshold is manipulated. The manipulation process is basically a deduction from the threshold based on the urgency value.

For each time step, the new threshold is less than the current threshold by σ*Fixed Value (Where σ is the urgency value between 0-1). Additionally, the threshold manipulation process resets the threshold when it receives that the action for the stimulus has been triggered. There are three behaviors for this process as the following:

- **Idle:** When the urgency signal is equal to zero and when the action is executed for the stimulus.
- **Manipulate the threshold:** when the action is not yet executed and the urgency signal has a value.
- **Reset the threshold:** When the action is executed, and the current threshold is lower than the initial threshold.

Process 3: Context Checking This process is receiving the stimulus states coming from the first process and sends an urgency signal to the threshold manipulation process. Additionally, it reads the salient hot point from the attention manager, analyzes the relevance of this point based on the given context, and finally sends action execution commands if it fulfills the action requirements. Finally, it updates the working memory with the action execution state. This process also receives a trigger command from the audio power trigger. Once received, it checks the relevance of the state, and based on that, it executes the urgency signal. Once the attention manager sends a hot point (when Γ exceeds the threshold). Once received, it starts the evaluation of this point in the task context. The evaluation is the relevance of the 3D projection of this point to the predefined working area in the environment. Knowing the 2D coordinates of the hot point received from the attention manager and the equation of the plane of the working area, we calculate the 3D location in the environment. Based on the defined task, the decision is made whether to do the action or not and which action to do based on the projected 3D location of the hot point. Once the action is executed, the urgency signal is set to zero and the action state is updated in the working memory. Following the assumption of ignoring the vertical component in the audio stimulation, we implement a function to force the vertical component of the 2D hot point to meet the location of the stimulation sources (defined in the context). This is done by estimating the vertical component given the current head altitude angle and the vertical field of vision. The robot identifies the stimulation source by calculating the distances between the projected 3D location and all the stimulation sources. The source corresponding to the minimum distance is the winning location.

The context is stored information related to the task and environment. This means that in this block, the task is defined with its requirement. The task is a defined action under a certain stimulation condition. The task-related information is information about the stimulation conditions, the starting level of confidence of the stimulation, the modulation rate which defines the urgency-accuracy trade-off, and finally the required action when the conditions are applied. On the other hand, the environment-related information in the action execution layer is a higher level of information. It includes the locations of the relevant stimulation sources, the working plane, and the action execution parameters. This information helps the robot to project the action from the 2D egocentric frame of the vision to the 3D world and execute it in a proper way. More information related to this section will be explained in the experimental setup section of the chapter.

To summarize this, the process starts with the audio power trigger. The trigger is perceived by the third process (Context checking). Based on the states and the task, the context-checking block executes the urgency signal to the threshold manipulation process. The threshold manipulation reads the current threshold, deducts from it based on the urgency value, and updates the threshold in the working memory and attention manager. Once the Γ value of the combined scene exceeds the threshold, the Hotpoint is streamed. It goes back to the context-checking process, which computes the 3D location of the stimulus, identifies the stimulus, executes the action, updates the urgency, and finally set the action execution state in the working memory. Once the threshold manipulation receives a zero urgency signal and the state of the action is executed, it resets the threshold to the initial value and updates the attention manager and the working memory.

4.2.3 Working Memory

The concept of working memory has emerged in psychology literature as a broad set of mechanisms that explain this accumulation of perceptual information over time. Psychology researchers have shown the relationship between attention and working memory [220, 60]. They have shown the irreplaceable role of working memory in solving cognitive problems by maintaining some essential information for certain tasks that involve monitoring the environment. Based on this information, we added a working memory element in our model to endow the robot with this ability. The working memory in our model maintains essential environmental and internal states for understanding the current scenario and for executing the correct action in the defined task. As shown in figure 4.2, the working memory block is bidirectionally connected to both the decision-making and perception components. In our implementation, we developed a state of working memory. It stores the states of the stimulation, action, and threshold (confidence) level to enable better interaction with the environment. The stimulation states define whether the stimulation is currently on or off and track it (for both the audio-visual scene and also for audio specifically). The reason for having a specific audio stimulation state is that it is a process that requires some time to be localized, unlike the visual modality. The audio stimulation state is set based on the audio trigger, while the combined audio-visual stimulation state is defined by the gamma value Γ of the scene with respect to the threshold level. If the gamma value exceeds the threshold, there is an on-stimulation. The attention manager block is responsible for maintaining the stimulation state. While the action states define whether the robot is executing the action or has finished the execution or still hasn't executed it for the current active stimulation. The decision-making controller maintains the state of the action execution as well as the confidence threshold as explained in the previous section. The attention manager and the decision-making blocks are recalling these states in their processes. The working memory block ensures a stable robotic behavior for attention, decision-making, and action execution cycle.

Another aspect of the working memory system is the habituation process. It is a perceptual stage necessary for the robot to memorize the specific conditions of the environment, as well as details about the human partner. Habituation is a well-studied process in psychology and neuroscience. It is the simplest form of learning [196]. It is defined as the process of learning how to filter out irrelevant stimulation and focus only on the important stimulation. [76] [248]. It is an important biological process for effective learning. In this work, we implement a simple form of habituation that allows the robot to learn the baseline sensorial characteristics of the environment and of the human partner in order to properly compensate during the task.

From the implementation point of view, we developed a habituation process for the cognitive architecture. The process can be represented as auto-calibration of the thresholds in the architectures. More specifically, the initial confidence threshold for the audio-visual scene which is the one the decision-making block is manipulating, and the threshold of the auditory power. This habituation (auto-calibration) process is executed after receiving a habituation signal that is sent to the decision-making block. This signal changes the current task to calculate some parameters from the scene in a defined time period. The parameters are the maximum, minimum, average, and standard deviation of the gamma value Γ and the audio power. The habituation also informs the process that the stimulation will be presented, and it is required to see the effect of this stimulation and memorize it. The effect of the stimulation is the changes in the gamma value when the stimulation is present. After the defined time period for the habituation process, the initial threshold of the confidence is set by the maximum Γ value during the habituation process, minus a fixed value as a sensitivity zone. The initial threshold value is one of the relevant details in the human-robot collaboration with the human partner. In particular, this threshold changes based on the visual environment, which includes the presence of the human subject, and the lighting condition. Similarly, the audio threshold is also calibrated this way. But since the auditory environment does not particularly change much with the existence of the human (of course if he doesn't speak), the habituation process has been executed once.

4.2.4 Action Execution

The action execution block receives commands from the decision-making block and then executes these commands by performing whole-body motor execution of a required action. The action is previously learned by the robot. The motor action execution is expecting an allocentric location in the working environment. By providing a reasonable assumption about the task, its context, and working area, we were able to define the attentive plane in a geometrical representation. Applying projection on this plane we estimate the allocentric representation of the required point. Based on the task, we assess the spatial relevance of this point and check if this point relies on the predefined working area of the current task. The implemented module for the action execution is called the attention action linker.

Attention Action Linker The attention action linker interprets the decision and executes the motor commands. The decision-making layer gives the command to the action execution layer with the result of the decision task. The linker also controls the motor action by enabling or by disabling it. The actions are predefined in the current task. In corresponding to the stimulation source there are two actions, the gaze action, and the point action. This part of the architecture is more task oriented. In this module, the response actions of the robot are defined based on the stimulus location. The main goal of putting this module in the architecture is to enable taking actions after finishing the perception process and making an attentional decision. In the Experimental part, we will talk about the Implemented actions for the defined task in the experiment.

4.2.5 Incremental Approach

To sum up, our main contribution in this chapter and the previous chapter (chapter 3) is the integration of perceptual processes, working memory and its rule in attention, time-variant decision-making, and finally the action execution into a complete cognitive architecture. Delving deeper into the details there are five main contributions as the following:

- **Audio Salient Based Allocentric Attention Representation**: adding new modules on the top of the audio Bayesian localization model to transfer the Bayesian map to a saliency map.
- **Audiovisual Integration**: by embedding the audio saliency map as another feature map in the linear combination of the PROVISION model.
- **Prior knowledge Integration Into the Audio Attention Component** to improve the localization abilities of the robot. (The priors are given explicitly in this version, however, it can be given as a top-down modulation knowledge from a higher perceptual process using vision or other cognitive modalities)
- **Computational Implementation of The Time-variant Decision Making (Threshold Manipulation)** which addresses the confidence-urgency trade-off in perceptual decisions.
- **Working Memory Integration in the Cognitive Architecture** to regulate the attention action cycle and track the state of the stimuli and the action.

The second two points are the main scientific novelty of this work. We introduce the time-variant decision-making process with the aid of working memory in the attentional decision task. This process is biologically inspired. It is embedded in an audio-visual cognitive architecture for attention and applied to the iCub robot to operate actively in real-time.

4.3 The Experiment

We test our three hypotheses mentioned in section 4.1.3 by performing a joint human-robot attentional task in an unstructured environment. The rationale behind the design of this experiment is the facilitation of the decision-making process evaluation, the performance of the system in different stimulation modes (audio-visual vs audio only), and finally, the comparison between human and robot performances. Figure 4.5 shows the experimental setup. The robot is facing the human participant. In between, there is a table that has the stimulation board and a keyboard in front of the human participant. The stimulation board is approximately centralized between the robot and the human with 57 centimeters of distance to both. The height of the chair where the participant sits is configured so that the human is

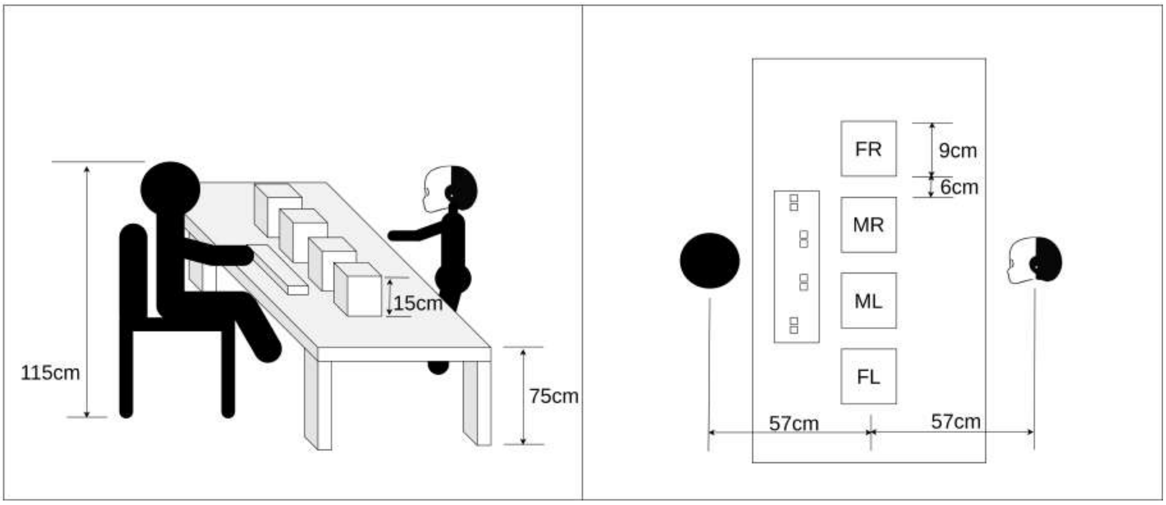


Figure 4.5 Experiment setup showing the positioning of the robot and the participant. Also, the four stimulation boxes and their locations. Far left "FL", Middle left "ML", Middle right "MR", and Far right "FR"

on the same level as the robot. This height places the stimulation board within an optimal location for the field of vision for both the robot and the participant.

4.3.1 Participants

We conducted the experiment with 21 healthy participants (female: 14, male: 9) aged between 26 and 43 years old, with an average age equal to 30.5 ± 4. All participants voluntarily participated and signed an ethical and informed consent approved by an ethical committee at San Martino Hospital in Genoa, Italy. All the participants work within the institution with no direct involvement in the research.

4.3.2 Stimulation

We built a stimulus setup that consists of four identical boxes. The boxes are placed horizontally on the same line. We noted the names of the boxes with respect to the robot's frame of reference: (FL) for the far left box, (ML) for the middle left box, (MR) for the middle right box, and (FR) for the far right box. Each box can produce both audio stimuli and visual stimuli. The visual stimuli are produced by a smart bulb. The smart bulb emits up to 800 luminous fluxes. We use red color with the maximum luminosity. The audio stimuli are produced by a three-watt Bluetooth speaker. Both the bulb and speakers are embedded inside the box. The top layer of each box has holes where the light and sound waves can propagate through, but that hides the smart bulb. The width of the box is 9 cm. The boxes are placed with a 15-centimeter separation distance (center to center). Therefore, the distance that separates the boxes is 6 cm. We placed the stimulation boxes in this configuration with the given spacing to make sure that all boxes are within the direct field of view (the view with a zero yaw angle for the face) of both the robot and the human participant. Additionally, we made the task more challenging by minimizing the distance between the boxes. As it is proven that human perception matches sound sources and visual sources for angles as large as 30 degrees apart [95]. we selected a long distance as half of 30 degrees and a short distance as one-fourth of these 30 degrees. This drove our choice for the configuration setup. We use a complex tone with a 1 kHz fundamental frequency and 3 harmonics for audio stimulation. The visual stimulus is a red light emitted from a smart bulb. The choice of the complex frequency and the red color is because of their high saliency compared to other colors for the vision, and simple tone for the audio. This was chosen to ease detection for both humans and robots.

4.3.3 Task description

The task for both the human and the robot is to identify the active stimulation box and react as quickly as possible. There are two types of activation for the stimulation boxes. The first type is audio-only stimuli and the second type is audio-visual stimuli. Only one box can be activated at a time. The stimuli are activated for a fixed time (10 seconds). The time between rounds is also fixed at 10 seconds. The experiment consists of 32 trials for each participant. The stimulation trials were distributed equally over the four boxes. So, each box was turned on 25% of all trials. Also, the stimulation types were distributed equally. 50 % of the trials were auditory-only and the other 50% are audio-visual. Each box was activated for 8 trials, 4 of them were audio-only and the other 4 were audio-visual. The sequence of trials and the type of stimulation were randomized but fixed across participants.

In the implementation section, we mentioned that the user defines the task for the robot and gives the system the required information for the task and its environment. Therefore, we defined the task on the top of the attention system. The task is to localize the stimulation from a set of defined sources located horizontally in front of the robot. After localizing the location, the robot should execute the gaze action (to look to the stimulation source) and point action (to point with the arms to the stimulation source). We provided the robot with environment-related info which is the working plane where the stimulation sources are located, and the working area on this plane. Additionally, we informed the robot that the stimulation sources are in that defined area in the space. Consequently, any localized stimulation within this area is considered as relevant to the task. If the localized stimulation is outside this area, then the robot ignores it as it is irrelevant stimulation. Extra environment information was added to the robot here, including the stimulation sources count and location. After localizing the 3D location of stimulation, the robot should identify the source of this stimulation from the defined set of sources. To sum up, the task is stimulation localization which is estimated in the decision-making layer. This task is divided into 2 stages, the first stage is localizing the stimulation within the 2D frame and the second stage is to check the relevance of this stimulation when the 2D location is projected into the 3D world. If it is relevant, then the robot will execute the action. The next section is describing the defined actions for the robot and also for the human participant.

4.3.4 Participants' Reaction and Robot's Reaction

We placed a keyboard in front of the human participant. On this keyboard, eight buttons were highlighted in four groups. Each group consisted of two side-by-side buttons. The human participants were requested to react as fast as possible by pressing any of the two buttons within the buttons group, which correlated to the activated stimulation box. We decided to use two buttons on the keyboard to increase the pressing area in order to simplify the action and minimize the execution time. For the robot, we defined two actions associated with each localized stimulation box. The first action is a pointing action using the arm, the hand, and the fingers while the second action is a gaze action using the head and the cameras (eyes) of the robot. For the right-side boxes, the robot will point to the selected box (FR, MR) using its right hand. Similarly, the left hand is used for the left side boxes (FL, ML). For the gaze action, movements in the head and the cameras are involved. The reaching action is a biological and human-like movement that recruits not only the entire upper body of the humanoid robot iCub but also the control of the head and gaze of the robot. The gaze action brings the fixation point (line of sight) on the target with optimal coordination of the 6 degrees of freedom of the head and eyes. Pointing with the index finger of the most opportune hand brings the robot to assume a new posture in less than two seconds. The coordination between head movement and upper body movement is designed in detail and makes the whole body movement look natural and human-like. It is possible that the human participant‘s attention is biased by this movement, but this is useful information in order to estimate the human-robot mutual influence in joint tasks.

4.3.5 Measurements and rounds

The robot and the human do the task together at the same time. Before the first trial for each subject, we introduced visual stimulation for the robot and the human. The robot performed the habituation process with the starting signal during this stimulation introduction period. Our first aim was to compare the performance of the robot versus the performance of the human participants in terms of both accuracy and reaction time. In general, we were also interested in measuring how much one participant influences the other in human-robot collaboration. In order to measure accuracy and reaction time for the human participant we recorded the pressed keys and their correspondence to the target as well as the reaction time. For the robot's accuracy and reaction time, we recorded the action execution commands of the robot and the internal triggering commands of these actions as relevant information about the timing and selected location. Additionally, we aimed to analyze all components of the decision-making processes. Thus, we recorded the threshold profile (indicating the urgency to act) as well as the integrated scene analysis which includes the Γ value (indicating confidence in the target localization process) during the whole trial. The second aim was to understand the behavior of the human participants considering the presence of the robot. Specifically, in this experiment, we focused on gaze behavior. We recorded the gaze data during the whole experiment using Tobii pro glasses. This data includes the 2D gaze location within the field of view of the camera and the gaze event (Fixation / Saccade). This is the main data from the eye tracker that we focused on. For better analysis, we developed a program to ensure synchronization between the eye tracker time stamp and the time stamp from our system. The idea of the program is to send a timestamp instance from our system to the Tobii pro glasses, and in the analysis stage, we map the timestamp of the eye tracker to our system's timestamp. The synchronization process ensures the transfer of the trials' information to the gaze data. The trials' information mainly includes the current state of the stimulation, the active box, the starting time of the trial, and the type of stimulation.

4.4 Results

We primarily focused on the assessment of the performance of the memory-based cognitive architecture for the attention task (jointly with the human). To perform an extensive evaluation of the system, we subdivided the analysis into two main sections. The first section is an analysis related exclusively to the performance of the cognitive architecture. This includes the evaluation of the whole system dynamics which is mainly the decision-making process and the overall performance (localization accuracy and reaction time) by comparing it with human performance.

The second part of the results is a detailed analysis of the gaze patterns. Given a thorough description of how the focus of attention was jointly redeployed, we focused our secondary analysis on the gaze patterns of both the robot and human participants. Such gaze behaviors are a direct result of attentional processing but more importantly tend to cause mutual influence between the robot and the human. Humans tend to look where their partner directs their gaze [63]. Also, it is an important component in joint attention [257]. So, the actions of the robot which are the gaze movement and pointing might influence the attention of the human toward a specific location. On the other hand, the gaze action of the human changes the visual features of the scene while the head moves. Consequently, this creates changes in the saliency map of the robot which might change its behavior, and this is what we want to analyze.

4.4.1 The performance analysis

The Memory Based Decision-Making Process

We evaluated the memory-based decision-making process to report how the cognitive architecture makes the decision to act, averaged across all trials. The process is based on working memory, the confidence measure (Γ value), and the confidence time-variant threshold (the threshold at which if confidence is reached, the agent will make a decision) as core factors of the decision-making process. The cognitive system makes the decision to act in presence of the event of crossing between the confidence measure and the threshold curve. Therefore, we analyzed the decision-making behavior to assess the effect of working memory as well as the performance of the confidence measure and the confidence threshold, which are core factors of the decision-making process.

Adding working memory allowed the robot to track the stimulation state of the trial (presence of a stimulation), and the state of his own action (whether the action is done, in progress, or not yet executed). This has a clear advantage with respect to other work done in the recent past [71] endowing the robot with the tracking capability and the understanding of the current state instead of direct control of the robot's behavior through a hard-coded state machine for a specific task. Once the robot executed an action for a certain stimulus, it could realize that the task is done and there is no need to execute the action again until the current stimulus stops. This represents its internal working memory of the active motor actions. When the stimulus stops the working memory is updated, allowing the robot to reset and wait for another stimulus.

Analyzing the stability of the system, the robot was successfully able to execute the action in the right time frame (after the stimulus turned on and before it turned off) in 95.8% of the trials. Comparing this result with the previous version presented in chapter 3 the behavior of the robot has been stabilized. Clarifying this point more, previously the robot was executing multiple actions to the same stimulus during the on time and also executing actions during the off time. The main added component is the decision-making mechanism. The decision-making controller with the aid of working memory stabilizes the action cycle and also allows the execution of the action based on meaningful environmental and internal states. This leads us to accept the first hypothesis **"Empowering the cognitive architecture with a memory-based decision-making process will stabilize the attentional performance."**.

Moving to the analysis of the confidence measure (Γ value) and the confidence threshold manipulation process, we present here the results by showing three figures as the following:

- Figure 4.6 is showing both the confidence measure Γ and the threshold for a trial to illustrate the behavior.
- Figure 4.7 is showing the confidence threshold profile averaged across all trials for both audio-only and audio-visual conditions.
- Figure 4.8 is showing the confidence measure Γ averaged across all trials for both audio-only and audio-visual conditions.

In this part, I will explain these three figures. Starting from the first one (Figure 4.6), it shows a single trial taken from one participant. Once the simulation starts, the threshold of confidence starts to decrease in time with a decreasing factor from the initial value (the parameter is specific to the participant, computed during calibration, and kept in memory by the system). The level of confidence indicated by the Γ function and the threshold profile progresses in time under their proper temporal dynamics until the Γ value and the threshold cross each other. At this point, the cognitive architecture makes a decision and acts, by consequently pointing to the estimated source of stimulation. Once the stimulation ends (after 10 seconds from its beginning) the system then resets the threshold to the initial value (the upper bound).

Moving to the second figure (Figure 4.7), it shows the average threshold profile with audio-only trials in blue and audio-visual trials in orange. The initial value of the threshold (the upper bound) is different for each participant. This is due to the habituation (Calibration)

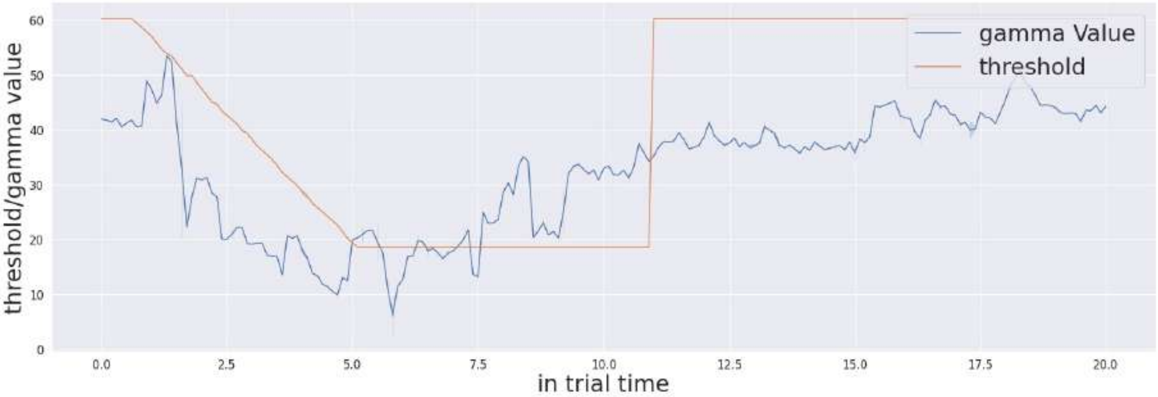


Figure 4.6 Gamma measure (Γ) and the threshold profiles in one of the trials across the whole trial time (20 seconds). The crossing occurs around the second 5

process, as the system memorizes a different initial value for the threshold for each participant. The process runs at the beginning of the experiment for each participant because this initial threshold is dependent on the visual features of the environment including the human participant in the field of view. The starting time of the threshold modulation process is based on detecting the existence of the stimulation. Thus, the exact starting time of the modulation signal is different from one trial to another. The confidence measure is incriminating in time which is due to the behavior of the Bayesian auditory model. Both the threshold modulation and the incremental value of confidence are defining the action execution time (the time when the confidence value crosses the threshold value). Therefore, the linear decrease of the threshold (manipulation process) explained in the first figure (Figure 4.6) became curved when the response is averaged across all trials. In the audio-visual condition, the threshold stops decreasing earlier. This is because the action is typically executed earlier due to the greater level of confidence in target localization. After the multi-sensory stimulation stops (experimentally fixed in time after 10 seconds from the beginning of the stimulation), the threshold resets again to the initial value. In this exact moment, in the audio-only trials, the threshold starts from a lower value. This reflects the lower confidence and consequently the longer response time to take a decision to act.

On the other hand, by looking at the third figure (Figure 4.8), we observe that the Γ function in audio-visual trials (orange curve) produces a spike almost instantaneously after the beginning of the stimulation. This is due to the visual saliency of the stimulation, which provides a strong, unique visual stimulation in the field of view. In the audio-only trials, the Γ function shows that the confidence decreases at the beginning as a causal effect of proactive sensing (the robot tries to eliminate the effect of the environment noise) and it

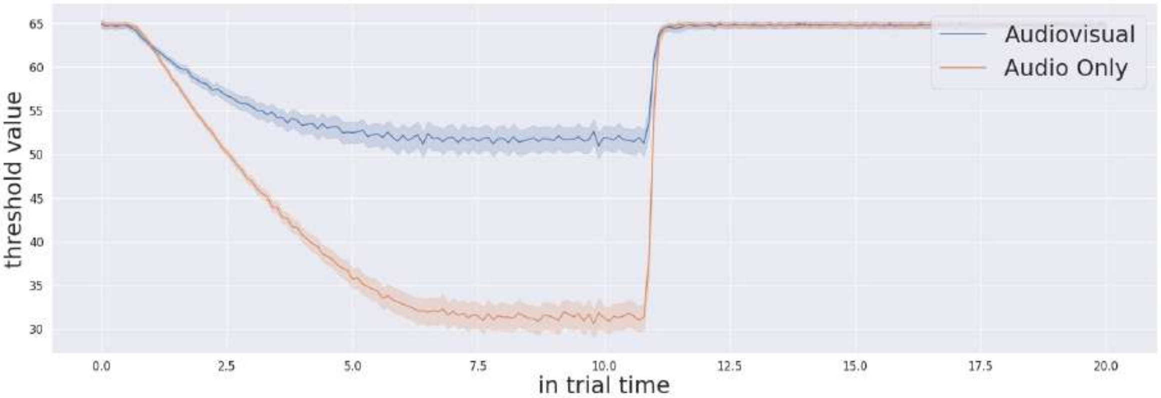


Figure 4.7 Confidence threshold profile across the whole trial time (20 seconds) for both audio-visual and audio-only trials

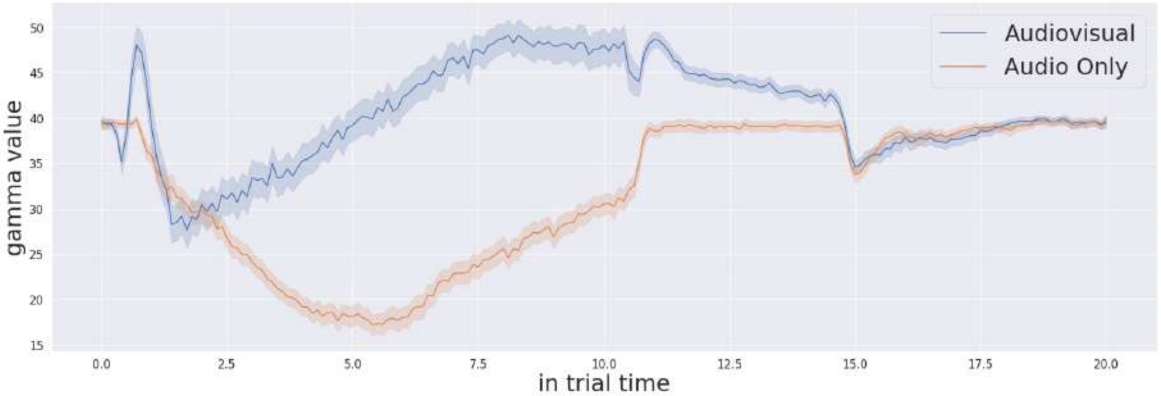


Figure 4.8 Confidence profile (gamma measure Γ) across the whole trial time (20 seconds) for both audio-visual and audio-only trials

starts to increase (after approximately 6 seconds in average) till the stimulation ends. When the threshold profile and the Γ measure cross on each other, the cognitive system makes a decision that triggers the action of pointing to the target stimulus.

The starting and stopping of the trial stimulation are autonomously detected by the system based on the audio power in the audio signals received by both the microphones as presented here in figure 4.9. The resetting of the threshold profile to the original value occurs after the end of stimulation is detected.

The overall performance (accuracy and Reaction time)

To assess the performance of the robot, we compared the attention system of the robot with human performance in response to the same multi-sensory stimulation and mutual sensorial influence. We analyzed the overall performance based on a) the reaction time

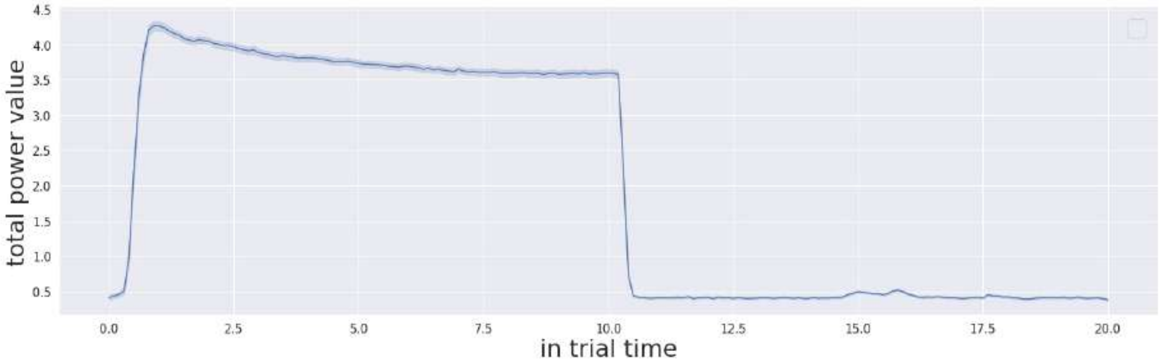


Figure 4.9 Audio power profile across the whole trial time (20 seconds) during all the trials

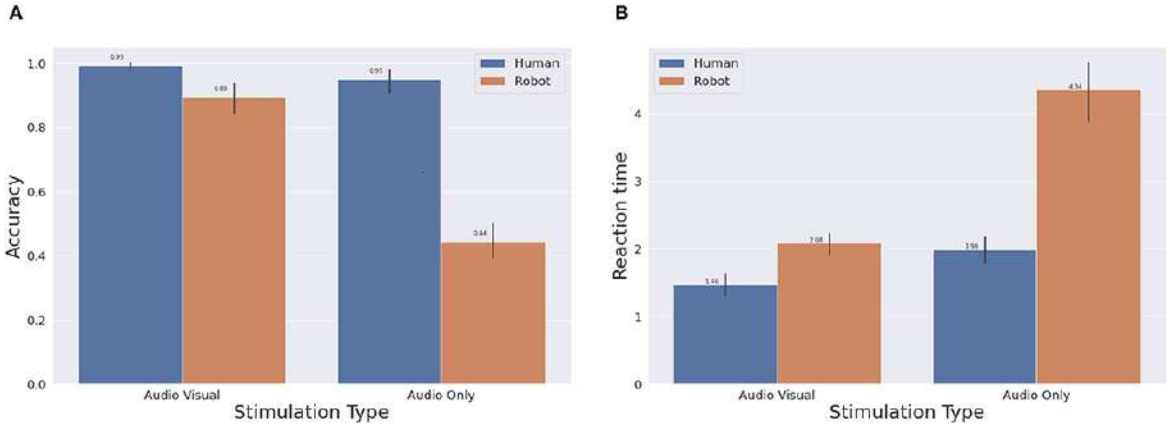


Figure 4.10 Overall performance across the different types of stimulation for both the robot and the human participants. It can be observed that the robot is less accurate and slower especially in the audio-only condition.

and b) accuracy as the primary source of evaluation. In particular, we characterized the performance based on the two stimulus typologies: audio-only stimulus and audio-visual stimulus. Figure 4.10 shows the measure of the reaction time and accuracy for both the robot and the human participants averaged across all the trials/participants. The bars in orange indicate the performance of the robot and the blue bars indicate the performance of the human participant. The participant and robot‘s choice is considered wrong if the identified box wasn’t the active box or if the action didn’t execute. Looking into the accuracy for each of the stimulus types separately, the robot records similar performance to the human in audio-visual attention tasks. The robot autonomously identified the source of the stimulation with 89% average accuracy. On the other hand, the robot performed with 43% average accuracy in the audio-only trials. The audio-only trials were more challenging for humans as well.

To assess performance, we performed multiple t-tests to compare the behavior of the human in the audio-visual task vs audio-only task, and similarly for the robot. The results of all the tests demonstrated significant differences as shown in Table 4.1:

Comparison	**t-value (40)**	**p-value**
Human audio-visual reaction time vs human audio-only reaction time	-3.7527	< 0.001
Human audio-visual accuracy vs audio only accuracy	2.1436	0.0382
Robot audio-visual reaction time vs robot audio-only reaction time	-9.6	< 0.001
Robot audio-visual accuracy vs robot audio only accuracy	12.2	< 0.001

Table 4.1 Comparison of audio-visual vs. audio-only performance accuracy and reaction time

So there are significant differences in both reaction time and accuracy between the audio-visual condition and audio-only condition for the human participants and also for the robot. The differences in the case of the robot were all significant (p < 0.001). As the average accuracy value for audio-visual is higher, and the reaction time is lower compared to the audio-only task (shown in Figure 4.10), we accept our second hypothesis that **"The stimulus localization accuracy and reaction time of the robot in the audio-visual task is better than in the audio only tasks"**.

We also performed t-tests to compare the performance (reaction time and accuracy) of the robot vs the performance of the human in the localization task. The statistical tests showed significant differences between both performances shown in Table 4.2:

Comparison	**t-value (40)**	**p-value**
Human audio-visual reaction time vs Robot audio-visual reaction time	-4.99	< 0.001
Human audio-visual accuracy vs Robot audio-visual accuracy	4.06	< 0.001
Human audio-only reaction time vs Robot audio-only reaction time	-9.7	< 0.001
Human audio-only accuracy vs Robot audio-only accuracy	14.9	< 0.001

Table 4.2 Comparison of human and robot performance (Accuracy and reaction time)

Thus, we reject our third hypothesis **The performance (accuracy and reaction time) of the robot will be as good as the performance of the human participants in localizing the stimulus**.

We did further statistical investigations using Wilcoxon signed ranked test [205] to test how different the performance of the robot was compared to the human. We found that the accuracy drop in the audio-visual condition is statistically less than 20% of the human

Table 4.3 Robot's Failure types percentages

	Failure		Percentage from total Failures
Type 1	Wrong Identification		89.4%
Type 2	No action (Wrong action type in previous trial)	60%	6.4%
	No action (low confidence)	40%	4.2%

accuracy. Also, the difference in reaction time of the robot in the audio-visual condition compared to the reaction time of the human is less than one second which is 70% of the increase in human reaction time. For the audio-only condition, the difference was much bigger than for the audio-visual condition. The differences in the audio-visual condition are comparable considering the complexity of the system and the processing speed of the machine. The audio-only condition is more complex compared to the audio-visual condition for both the human and the robot. However, the complexity of the audio-only localization task does not entirely explain the considerable gap. To understand the reasons for this performance drop, we more thoroughly investigated the conditions of wrong actions. The results are shown in Table 4.3. There are two conditions in which we consider the behavior of the robot to be worse. The first condition is when the action is executed but the identification of the active box was wrong and is annotated with "wrong identification". The second condition occurs if the action is never executed during the time of the trial and we annotate this behavior as "no action". For humans, all the wrong action trials were due to wrong identification. For the robot, the first condition occurred most of the time (89% of the total failures). On the other hand, there were two causes for no action failures. The first cause is when the robot executes an action in the off time of the stimulation due to some confusion from visual features in the scene. More specifically, it was observed that for some participants the robot got confused from the hand of the participant, indicating once again how mutual influence impacts attentive tasks. The participants' hands worked as visual stimulation and the robot identified the closest box to the hand as a source of stimulation during the off time. If the robot executed an action during the off time, the robot does not reset the exception event before the end of stimulation of the next trial. The consequence of this is a (no action) failure for the trial next to the off time when the robot executed the action. This actually happened very few times (15 times) across all trials, which consists of 6% of the total failures. This is 60% of the second type of robot failure (No action failure). The remaining 40% of no-action failures are due to low confidence levels. The robot did not execute an action a few times because the confidence value (Γ value) never reached the threshold during the time of the trial. This type of failure only forms 4% of the total failures.

Based on these analyses, the major cause of failure is wrong identification. Therefore, it is also important to analyze in detail the attentive process in time. More specifically, the audio components need to be analyzed, because the difference in performance lies in the temporal response of the attention system. So, in the next section, we analyze the temporal responses of the audio probabilities, which are the base of the localization process during the audio-only condition.

Detailed analysis of the audio-only trials (probability profile)

Since the behavior of the decision-making process does not show erroneous behavior, but instead the decision is made in the right time frame with a reasonable level of confidence, we believe that the reason for the worse performance in audio-only trials is to be found in the localization process. As shown by a more detailed analysis for audio-only trials, the localization process is based on the level of confidence that each box is the target, in other words, the probability that each one of the four locations is the target. Such probability changes over time for each potential location of a stimulation source. In the audio-only condition, the probability profile is extracted from the Bayesian map, which is the output of the audio localization system. The temporally detailed analysis of the probability profile is carried out during the 20-second time frame of the trials. During the first 10 seconds, the auditory stimulation is generated by the target box only.

Figures 4.11 and 4.12 show the probability profiles for the 4 locations of the stimulation sources when the active box is the far left one and the middle left one respectively. The response is averaged across all trials. The first relevant point of these figures is that the shape of the curves is similar for the boxes located on the same side, independent of the location of the source of stimulation. In other words, the probability profile over time of the far right is similar to the one middle right and similarly, the probability profile of the far left is similar to the middle left. Such results indicate that there are differences in the time progression of the probability profile between the left and right boxes from the location where the robot is standing. The such difference has an impact on the localization of the sound target since the certainty of sound location changes over time differences between the left and right boxes. Similar difficulty from one side over the other was actually reported by most of the participants. Another aspect that might have an impact on the localization of the source of sound is that the probability profile of the sound sources from the same side evolves similarly. This makes the discrimination task complex for the robot, but also for the human participant. It was challenging for them to identify which box between the 2 boxes on the same side is the stimulation source in audio-only trials. The similarity between a human-robot participant on

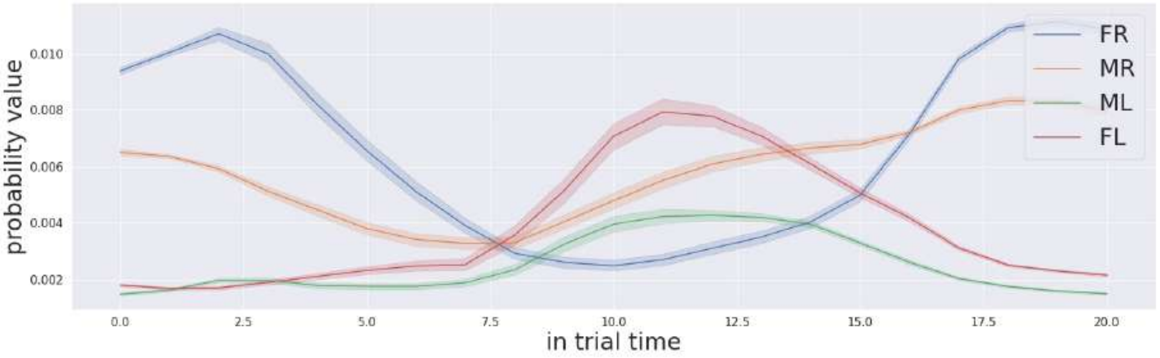


Figure 4.11 Audio probabilities profile for the trials that the far left box (FL in red colour) was activated

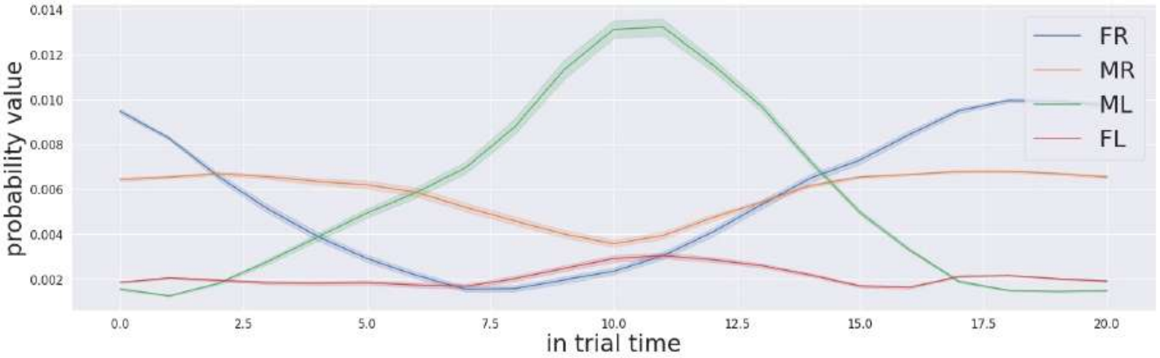


Figure 4.12 Audio probabilities profile for the trials that the middle left box (ML in green color) was activated

the same side during sound discrimination suggests that the Bayesian modeling implemented in the cognitive architecture shares some similarities with human behavior.

Another relevant point relates to the temporal profile of the probabilities for the different salient locations. The probability corresponding to the right location increases with time as long as the stimulus is active, (in the first 10 seconds) which is the right and required behavior. However, the probabilities of corresponding matches between the source of sound and different locations do not always start from zero and have equal values. This indicates that before the activation of the stimulation, the localization system believes that one location is more likely to produce sound than another location. Each probability goes to an initial value that is not equal to zero and also not equal to other locations' probabilities. Our speculation explains the presence of these two phenomena as the result of acoustic noise in the environment. The acoustic noise equally affects the performance of the robot and of the human participant. It would be wise to remove the constant acoustic noise in the environment to eliminate its effect on the Bayesian map probabilities first, and then integrate evidence from the actual stimulation over time.

The final consideration regards the time the system requires to make the right decision. From both graphs, we observe that it takes in approximately 7.5 seconds for the far left box to be the box with the highest probability and 6.5 seconds for the middle left box. For the boxes located on the right side of the robot, the value for the middle right is similar and is approximately 7 seconds. For the far left box, the system struggles due to the noise, the probability for the far left never reaches the maximum when the box was activated within the on-time frame (10 seconds). The decision-making process is tuned with some parameters to react faster than the required time. So the average reaction time of the robot for audio-only stimulus was measured to be around 4.34 seconds (STD: 1 second), definitely faster than the time necessary for the temporal probability profile to converge on the correct stimulation. Thus, we note that the attentive system can localize the target with a higher accuracy if the decision-making process is allowed a longer reaction time. However given enough time, the auditory localization process is always correct and the probability of the correct target always exceeds the probability of the others. For example, the audio probability profile for the far left box is the highest after 7.5 seconds. For the middle left box the audio probability profile for the middle left target is the highest after 6.5 seconds. Such fine refinement is actually doable in the cognitive architecture proposed since by adjusting the tuning parameter we can refine the decision-making process and adjust the decreasing rate of the threshold.

In conclusion, we assessed that the task results are also difficult for the human participants according to an interview in the debriefing phase of the experiment. Another relevant observation in regard to the numerous comments of many participants indicates the change in the auditory landscape as the most meaningful cue to localize the target. The suggestion convinced us to look at the change rate of the confidence level for the different possible targets. In figure 4.13 and 4.14 we show the change rate of the confidence probabilities for the four locations for the trial respectively when the target is FL and ML. We noticed that the attentive system can localize the target correctly in a shorter time if the decision-making process analyzes the change rate of the confidence probability instead of the confidence probability. For example, for the target in FL (see 4.13) the correct detection of the target can occur as early as approximately 3.0 seconds, and for the target in ML (see 4.14) the correction detection the target can occur as early as at approximately 2.5 seconds.

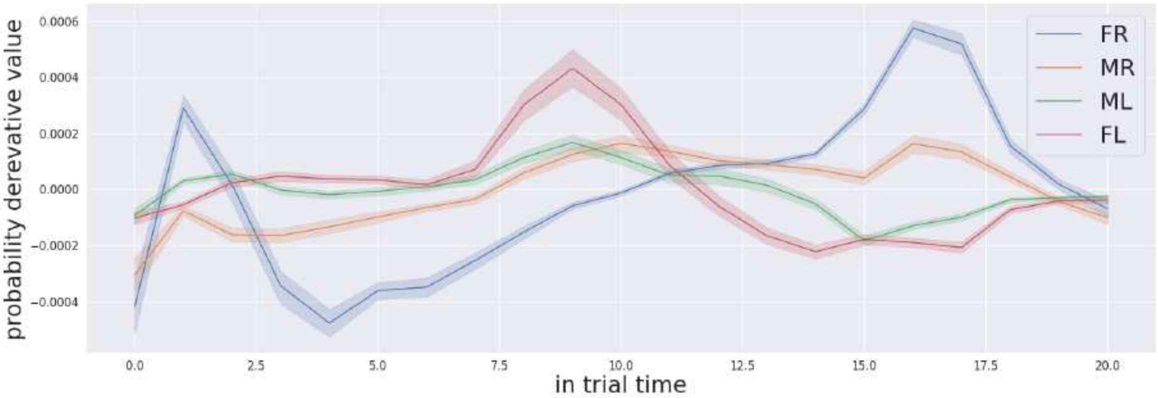


Figure 4.13 Derivative of the audio probabilities profile for the trials that the far left box (FL in red color) was activated

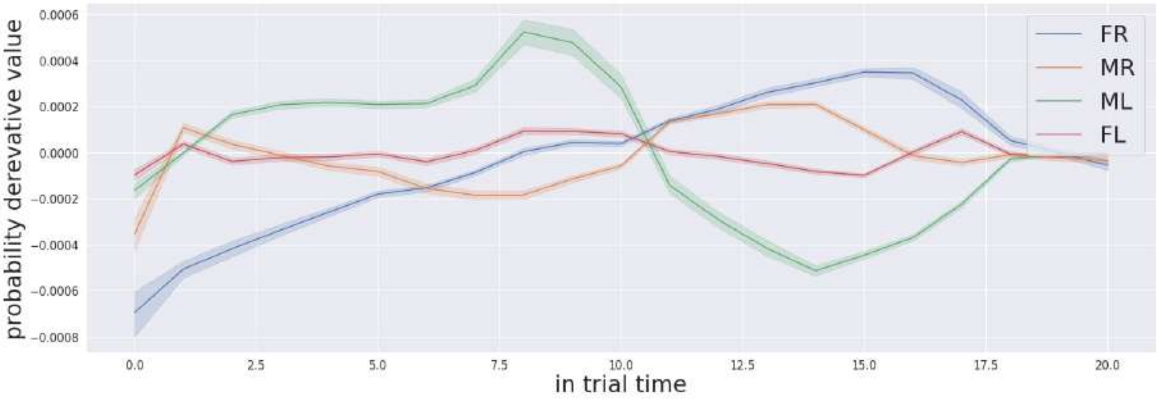


Figure 4.14 The derivative of the audio probabilities profile for the trials that the middle left box (ML in green color) was activated

4.4.2 The behavioral gaze analysis of the human and the robot

The behavioral analysis of the human participants gives us relevant insight into the mutual influence between the two partners. The behavioral analysis relies on data from the eye tracker. We were able to record the gaze data of the human participants. The gaze data is the 2D location of the gaze and the gaze event. The gaze events can be one of two types: fixation and saccade. We aimed to count the fixation events on the stimulation boxes and also on the robot's face during each trial. So, we had to define where the 2D location is projected in the 3D world. We are interested in 5 regions (the 4 stimulation boxes, the robot's head, and other areas). The eye tracker gives the 2D location of the gaze in the camera frame, which changes when the participant moves their head with respect to the world. In order to cluster the fixation events based on the 2D location into 6 clusters, we had to transfer the 2D location from the camera moving frame to a global fixed frame. We achieved this by extracting a reference point in the scene that always exists and then we track this point. This point works as a reference point and all interested regions are defined with respect to this point.

From the 21 subjects of the experiment, we could extract the gaze data perfectly from all 12 of them. Three subjects were moving their head very rapidly, and due to this, the process of extracting the reference was not accurate enough. The gaze data of 5 participants weren't accurate enough to be considered because the eye tracker failed to calibrate their eyes. So in this section, we only consider the data of the 12 subjects for which the calibration was accurate and the reference extraction process was sufficient. The robot's behavior in this experiment consists of its actions, which are the gaze movement toward the selected box and the pointing action with the arm. Fig 4.15 is showing the fixation distribution in trials. It is divided into 4 panels based on the location of the simulation. (FL, ML, MR, and FR for top left, top right, bottom left, and bottom right panel respectively). The y-axis shows the fixation counts. The x-axis here is the five defined regions of interest (4 stimulation boxes and the robot's head). We also categorize it based on the stimulation type: audio only in blue and audio-visual in orange. Similarly, Fig 4.16 shows the gaze of the robot. The robot only does one fixation event during each trial, which is the action of the task. So, the graph also represents the action distribution of the robot. The fixation counts on the active stimulation box are marked with a red rectangle surrounding the bars of this location in each of the panels for both the robot and the human participants. We divide our findings into two parts. The first part is for audio-only trials and the second part is for audio-visual trials.

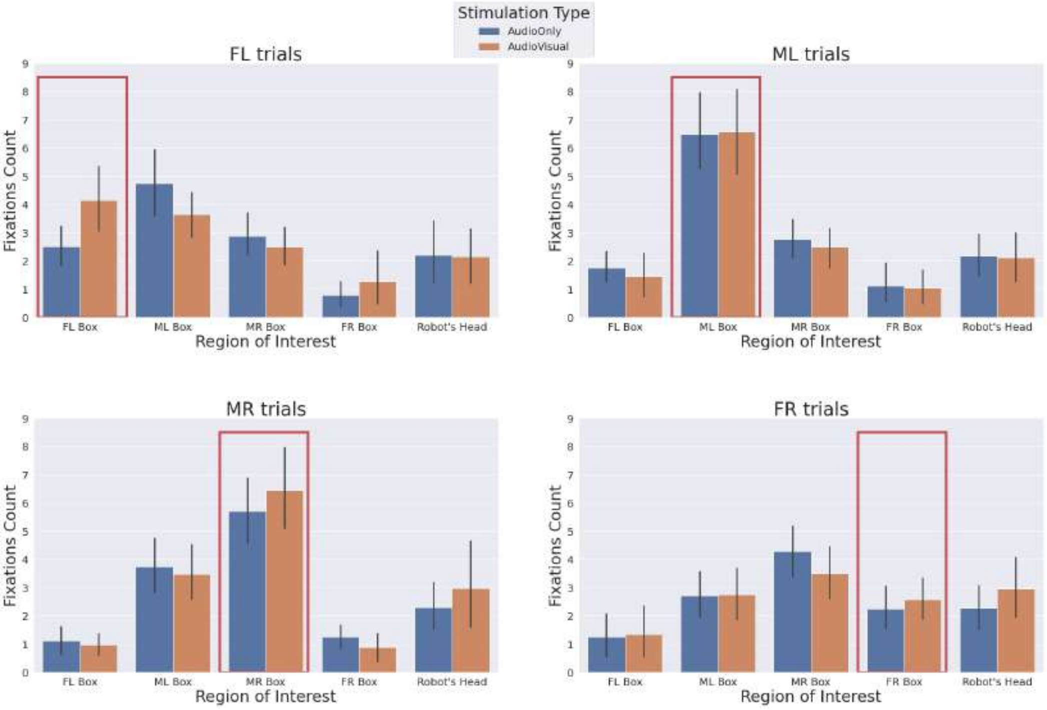


Figure 4.15 The gaze behavior of the human

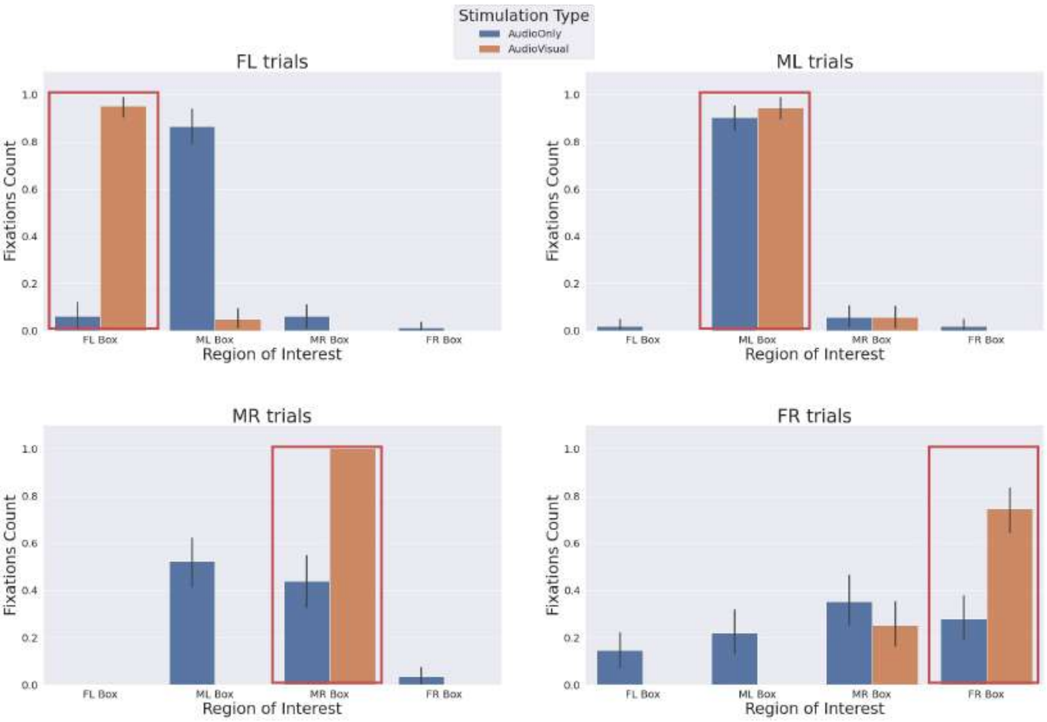


Figure 4.16 The gaze behavior of the robot

The first observed information is that in audio-visual trials the participants do fixation events on the active stimulation box more than other boxes in FL, ML, and MR trials. But in trials during which the FR box was active, the participants do more fixation events on the MR box on average. This drives us toward the second observation. Looking into the robot's gaze behavior, we found that in the FR trials, the robot was confused toward the MR box and sometimes performed gaze actions toward the MR box instead of FR. The next three observations are in the audio-only trials. In the FL trials, the robot mostly was driven toward the ML box. This records the highest average in comparison with the other boxes. Similarly, the participants also do more fixation events on the ML box, even more, the correct active box which is FL. The second observation in audio-only trials is in the ML trials. In these trials, both the robot and the participants do fixation events on the correct box more than on other boxes. Thirdly, in MR trials the confusion of the robot was between the right box (MR) and the ML box. But it is less than the confusion in FL trials. On the other hand, the participants' gaze record the highest count on the right box (MR) and the second highest is the ML box. Finally, it is clearly shown that the participants also spend time looking to the robot's head in all trials for all conditions.

4.5 Discussion & Conclusion

Joint attention is a fundamental component for better collaboration in real-world scenarios, such as in industrial environments where the robot and the human worker have to be aware of the products being manufactured (indicated by machinery through visual and audio features). They will be able to coordinate their actions and activities when initiated through their joint attention directed to the same target. The proposed biologically inspired cognitive framework, based on a multi-sensory attention system and supported by memory, constitutes the computational model used to evaluate emergent joint attention between the human participant and the artificial agent. The study had three main hypotheses. H1-Memory-based Decision-Making process: Empowering the cognitive architecture with a memory-based decision-making process will stabilize the attentional performance, H2-Audio-visual vs Audio-only: The stimulus localization accuracy and reaction time of the robot in the audiovisual condition will perform better than in the audio only condition. H3-Robot performance: The performance of the robot will be as good as the performance of human participants. To answer the hypothesis we designed a multi-sensory task and presented the task to the human participant and the robot. The setup includes stimulation boxes, which are a general model for real-world applications. Thus, we were able to compare the performance of the robot with the performance of a human participant in the same task which is an important aspect of defining the quality of the interaction. The comparison focuses on the assessment of both agents in terms of the execution of the same localization task with the same response time. The rationale behind the co-assessment of both the participants is that we intend to assess the performance of the robot and the human to measure how much they can coordinate in the joint task and to also measure the mutual influence between the robot and the participant.

The statistical analyses resulted in accepting the first two hypotheses (**H1-Memory-based decision-making process: Empowering the cognitive architecture with a memory-based decision-making process will stabilize the attentional performance** and **H2-Audio-visual vs Audio-only: The stimulus localization accuracy and reaction time of the robot in the audiovisual condition will perform better than in the audio-only condition**) and rejecting the last one **H3-Robot performance: The performance (accuracy and reaction time) of the robot will be as good as the performance of the human participants in localizing the stimulus**.

However, further statistical analyses showed that the performance of the robot in the audio-visual condition is comparable, as the accuracy drop was less than 20% of the human accuracy and the reaction time differences were less than one second which is less than 170% of the human reaction time. These values are acceptable considering the machine processing speed of such complex computational processes. Indeed the cognitive system is less reactive in audio-only stimulation and only partially influenced in the different internal processes by the presence of the human partner. Although the audio-only condition is in general a challenge for both the human and the robot participant, the analysis showed that the main cause of the performance drop in the audio-only condition is the false audio localization, which is caused by the acoustic egocentric noise. The main source of the egocentric is the fan on the back side of the iCub head. It disturbs the auditory localization system and moves the localization towards it which is further translated in the front field. The front-back confusion is one of the characteristics of the interaural time difference which is the method the Bayesian auditory localization model is based on.

Furthermore, we performed a more detailed analysis of the cognitive processes, and we realized that the decision-making process is robustly designed to swiftly guide the system to make a decision with excessively fast temporal dynamics. On the contrary, the auditory attention system requires longer time periods to make the Bayesian network converge, and thus localize the auditory target. Whereas the auditory localization process is correct in inferring the location, also in presence of environmental noise (typical in robotic applications), the temporal dynamics of the system require longer periods for the processing of the auditory stimulation. However, the specific inefficiency is of simple resolution for two reasons that we intend to verify in future work. First, the specific modular structure of the developed cognitive architecture and its parametric configuration is designed to allow for fast re-adaptation of the decision-making process. As one possibility, by reducing the urgency to act parameters in the decision-making process, we can allow more time for the Bayesian network to converge, and consequently, we can guarantee improved accuracy. However, although the specific solution improves the accuracy it does not guarantee a faster reaction time. Secondly, thanks to the margin for faster response during auditory localization, the process allows us to provide more auditory evidence for Bayesian integration in the same time interval. Faster processing of auditory stimulation is expected to improve the reaction time of the auditory localization system and make it more similar to the reaction time of human participants.

Undoubtedly, the temporal dynamics of how auditory evidence is integrated is a very important aspect. We noticed in human participants that changes in the auditory landscape are more meaningful for target localization than a static auditory landscape. The same process based on changes in the Bayesian network facilitates the process of inference over the stimuli localization. The importance of relative changes in the auditory landscape, together with the importance of proactively creating such changes in the auditory landscape (self-programmed head movements) is a promising area of study, and we are planning to exploit it further in future work. Nevertheless, even without these improvements, the cognitive architecture has been demonstrated to be effective, and it shows a natural and robust joint attentive behavior for Human-Robot interactive tasks. Furthermore, for a thorough understanding of behavior related to the mutual presence and its mutual influence, we also analyzed the gaze behavior of the human participants. The results showed that in the conditions in which the robot confused the location of the active box, the human participants tended to do more fixation events on the wrong box, suggested by the wrong behavior of the robot. Also, the participants spend time looking at the head of the robot during the experiment, which shows how the human participant and the robot mutually influence each other in similar interactive tasks. This brings us to conclude that the behavior of the robot may reinforce the gaze of the human toward the robot‘s chosen box. This is reinforced by the robot’s behavior which is both built on the directed gaze and the pointing actions. In the future, we intend to investigate this aspect further with more statistical evidence, and we intend to know whether this hypothesis of mutual reinforcement is confirmed and what exactly drives it: whether the gaze or the pointing or a combination of both have a stronger effect on the human partner.

Finally, we believe that the proposed system paves the way to human-robot collaboration since coordinated joint attention is proven to facilitate coordination between the interacting parts. Such an optimal mechanism of coordination is considered one of the main facilitation mechanisms in multi-partner interaction tasks. We also showed that the robot affects the gaze behavior of the participants. Furthermore, with this cognitive architecture, we demonstrate the importance of implementing a complete cognitive architecture (including working memory) in order to attend to salient targets in the environments as humans do. By sharing the same attentional focus redeployment mechanism with the human partner we provide effective joint attention that essentially emerges from environmental stimulation and reinforces natural human-robot collaboration.

With this work, we conclude this part of the thesis and answered the first research question **[*RQ1:*] How can we integrate state-of-the-art vision and audio models to allow the robot to jointly attend to the environment with a human partner? Is the behavior of the robot effectively received by the human partner? and What is the mutual influence between the robot and the human partner during the interaction using this model?)**. The next part of this thesis will go beyond the attentional mechanisms and explore a higher level of perception activities to endow the robot with human-like environmental proactive understanding and action coordination skills.

Part III

Proactive Audio-Visual Cognitive Architecture For Perception

Chapter 5

Main Implementation for The Audio-Visual Architecture for Perception

> The tendency of our perceptions is to emphasize increasingly the objective elements in an impression, unless we have some special reason, as artists have, for doing the opposite
>
> **Bertrand Russell**

The previous part of the thesis focused on addressing an attentional cognitive architecture for the robot to perform a joint attention task with a human partner and the mutual influence between them. The first part addressed the first research question of the thesis. Going one step further in the research, this part of the thesis is focusing on the implementation of a cognitive architecture for a human-like perception. With the target of endowing the robot with the skill of common representation of the environment with the human. The part will address this through three chapters (this chapter and the following two chapters). This chapter is addressing the main implementation of the architecture, followed by a chapter that explains the research to enable proactive behavior, and the last chapter of this part of the thesis will address how we applied this architecture in different robotic platforms and in a real-world scenario that requires high-level action coordination.

5.1 Background

In daily activities, humans are continuously stimulated in various senses. At any given time, the environment is perceived through multiple sensory modalities. The human brain is capable of integrating or segregating this sensory information and building an understanding of the current scene. For a situation where multiple stimulation sources are existing (e.g. two guitarists on a stage and one of them is playing guitar), the brain must explore and analyze the sensory input and then identify the playing guitarist correctly. Although these perceptual tasks seem trivial, it is computationally complex. These complexities are due to the differences between the representation of sensory information, and perceptual inference. The representation of each sensory information has its frame of reference as well as processing speed. For these reasons, cognitive studies proposed that the brain must adjust the information from different sensories to a unified metric before integration [9, 191, 51]. Following the adjustment, the inference of the information takes place based on the statistical probabilities of the information coming from each modality and combined into a multisensory representation [34, 48]. From a computational perspective, it is widely accepted that the inference mechanism is following the Bayesian strategy that leads to an optimal integration [53, 57]. In detail, the integration is performed as a weighted sum using the perceptual confidence (inverse of the sensory variance of the information coming from each modality) as the weights for the integration[117].

Another important point in multisensory is the cross-modal interaction. It refers to how sensory modality affects (influences) other modalities. This effect occurs in different perceptual levels including attention [50] and sensory integration [48]. One of these effects is due to prior knowledge and expectations. It is one of the important factors in cross-modal interaction and integration that has been recently shown in different studies [52, 36]. Prior knowledge refers to concluded information from the previous scenes that someone experienced. In the context of sensory integration, it is particularly how different sensory features are integrated together. Prior knowledge is a supporting factor in learning and adaptation activities that the brain does [68].

The research presented in this chapter aims to build a cognitive architecture (based on the previously presented work in part II of the thesis) for the audio-visual robotic perception that adopts the discussed multisensory perceptual processes (Bayesian multisensory inference, cross-modal interaction, and prior knowledge) and applies it to operate in real-time. Thus, we endow the robot with a human-like multisensory representation. Looking at the literature, some multisensory cognitive architectures and systems for robots have been proposed. For example, [246, 111, 33, 7, 172] proposed audio-visual systems for speaker detection and localization. [67] proposed an audio-visual system for navigation, Although their system uses both modalities, it doesn't do sensory inference which is one approach in the applications that doesn't require high accuracy in mapping the sources of stimulus. Also, a recent neural network model (AVOT) was proposed for audio-visual object tracking [253] that has been applied to videos (not in real-time).

So far, the previously developed audio-visual systems are designed to solve a specific task and use basic inference mechanisms. Additionally, none of these models address the cross-interaction between modalities [120] and some of them are applied in the real robot while others are not. Having a unified framework that does the biological inference processing with the use of hierarchical integration of priors is the challenge as mentioned in a recent review [129].

We propose a novel biologically inspired architecture for robot perception that aims to address these challenges which are, the inference mechanism, cross-modal interaction, applying it in real-time, and finally the generalized approach that is not task specific. In this chapter, I explain the main development of the architecture which is targeting the common representation skill category of shared perception (the second research question *RQ2*). The following section is giving a brief about the biological inspiration for my architecture.

5.1.1 Biological Inspiration

What and Where Pathways Ungerleider and Mishkin proposed a model for the visual pathways in the brain in the 80s [155, 242]. The idea has been well-established and proven throughout the years. It describes the processing of visual information in the brain. Which is segregated into two different specialized pathways. The ventral pathway, which is specialized in recognition (what), and the dorsal pathway which is specialized in localization (where). More recently, the two pathways for auditory processing were proposed and evidence was found in humans and other animals [140, 102, 203]. The dual pathway theory suggests that the what and where pathways are parallelly running and interacting with each other to create an overall perception of the world (Visual or Auditory) [123, 131]. Although there is some criticism of the dual pathway theory, it is widely accepted in the research community and supported by many studies. The other point that is related to the dual pathways is who then these pathways are integrated and create a perceptual object. The perceptual object (auditory or visual) is a distinct unit that the brain integrates from the dual pathways which is the result of the scene analysis [74, 155].

Perceptual Inference Perceptual inference is the process of interpreting sensory information and combining them together to form a representation of the environment. The inference processing is not just mapping the knowledge that comes from the senses but it also involves a reasoning process that integrated prior experiences as recently proposed [36]. For example, when a sound is heard, the brain interprets what is beyond this sound by recalling the previous knowledge about this sound which is formed previously. Perceptual inference is a key aspect of perception as it allows a complete understanding of the environment using multiple sensory channels. Bayesian probability theory [116] explains the inference processing in the brain and how it combines the information from different sensories as well as the prior knowledge in a probabilistic way. This is done by computing the evidence from the senses and integrating the priors using Bayes' rule. Sensory integration decides whether two signals from different senses are emitted from the same source or two independent sources. This integration is done through the estimation of the spatiotemporal properties of the signals as well as the priors (reasoning based on previous knowledge). Due to this strategy, different senses are affecting each other which is called cross-modal interaction. This is also associated with the cross-binding problem [211]. The cross-binding problem is defined as the difficulty of binding or integrating the right cues from different modalities together as a multimodal object in a certain situation. The situation can occur due to the processing and cognitive limitations of the brain. It can also happen due to damage in the brain that stops or limits some of the brain processes.

Scene understanding and spatial working memory Humans have a remarkable ability to perceive and understand complex scenes in the real world. This understanding of the environment is guided by spatial working memory. It refers to the ability to store information that is related to the spatial scene. Evidences showed this relations in attention [10] and

Spatial memory has a profound role in perception and many researchers in neuroscience showed this [229, 42]. Recent evidence showed that some cross-modal association processing is storing the focus of attention in working spatial memory [159]. So spatial working memory has an important role in cross-modal association. This inspired me to model this processing in my architecture. (Check [194] for a review on the latest research on multisensory working memory)

5.2 The Design and The Implementation of The Proposed Architecture

Figure 5.1 shows the overall structure of our proposed cognitive architecture. Our contribution is the design and implementation of the architecture. It consists of different kinds of blocks (shown in different colors) based on the type of processing the block performs. The complexity of the information increases from bottom to up. In architecture, there are two streams (vision and audio) that are integrated at a higher level. Each stream has attention, what pathway, where pathway, and a manager. Following this stage, information from each of the managers is sent to the short-term memory block. Also, the managers send commands to the action block. The multisensory manager governs the integration process with the aid of the long-term and short-term memory blocks. Also, it sends commands to the Action block and finally gives an output which is a set of audiovisual objects in the spatial space each time frame.

In the next parts of this section, we will explain in detail the implementation of the modules within each of these blocks and the information flow. Figure 5.2 shows the architecture in detail (modules level and their connections).

5.2.1 Audio and Visual Perception

The unisensory stage for both audio and visual streams consists of four main blocks as the following:

- Attention
- What Pathway
- Where Pathway
- Manager / Decision Maker

In this part of the chapter, each of these processes will be explained in detail for both audio and visual streams.

Visual Attention This architecture is an extension of the previous architectures with some modifications in the integration process. Therefore, the visual attention component in the architecture is still the PROVISION model. However, we added in this version a new module which is the **Inhibition of Return (IOR)** module.

The inhibition of return module suppresses the already attended zones/objects on the scene. This process allows the robot to shift its attention toward new spatial zones/objects. In this module, I defined the suppression zone as a circle with a gaussian faded distribution. The IOR process is a well-known and studied phenomenon in visual attention. It refers to the process of reducing/suppressing the priority or importance of a stimulus in a spatial location that was already observed (attended). This process acts as an orientational attention

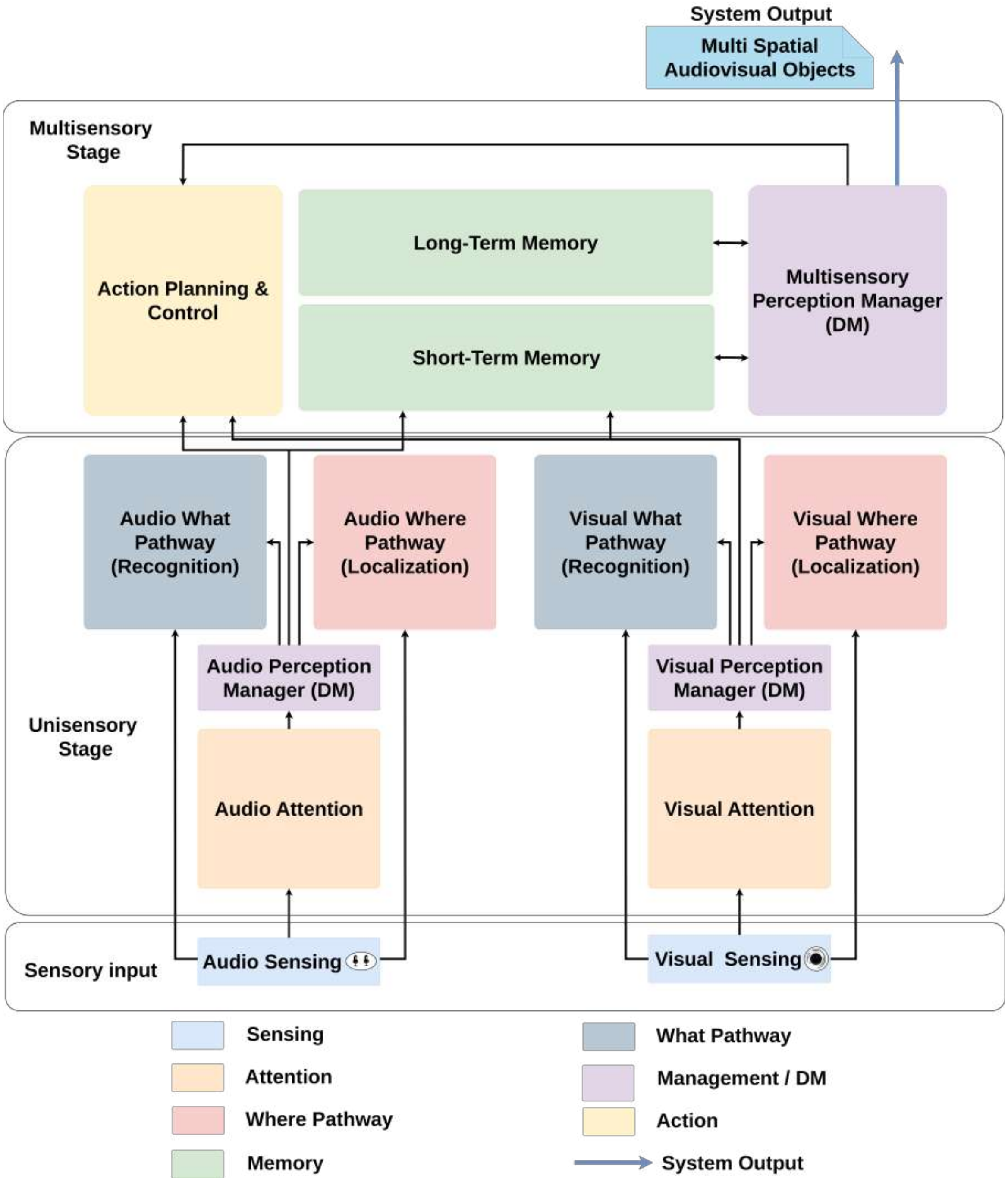

System Output
Multi Spatial Audiovisual Objects
Multisensory Stage
Action Planning & Control
Long-Term Memory
Short-Term Memory
Multisensory Perception Manager (DM)
Audio What Pathway (Recognition)
Audio Where Pathway (Localization)
Visual What Pathway (Recognition)
Visual Where Pathway (Localization)
Audio Perception Manager (DM)
Visual Perception Manager (DM)
Unisensory Stage
Audio Attention
Visual Attention
Sensory input
Audio Sensing
Visual Sensing
Sensing
Attention
Where Pathway
Memory
What Pathway
Management / DM
Action
System Output

Figure 5.1 The Overall Structure of The General Audio-visual Proposed Perception Architecture.

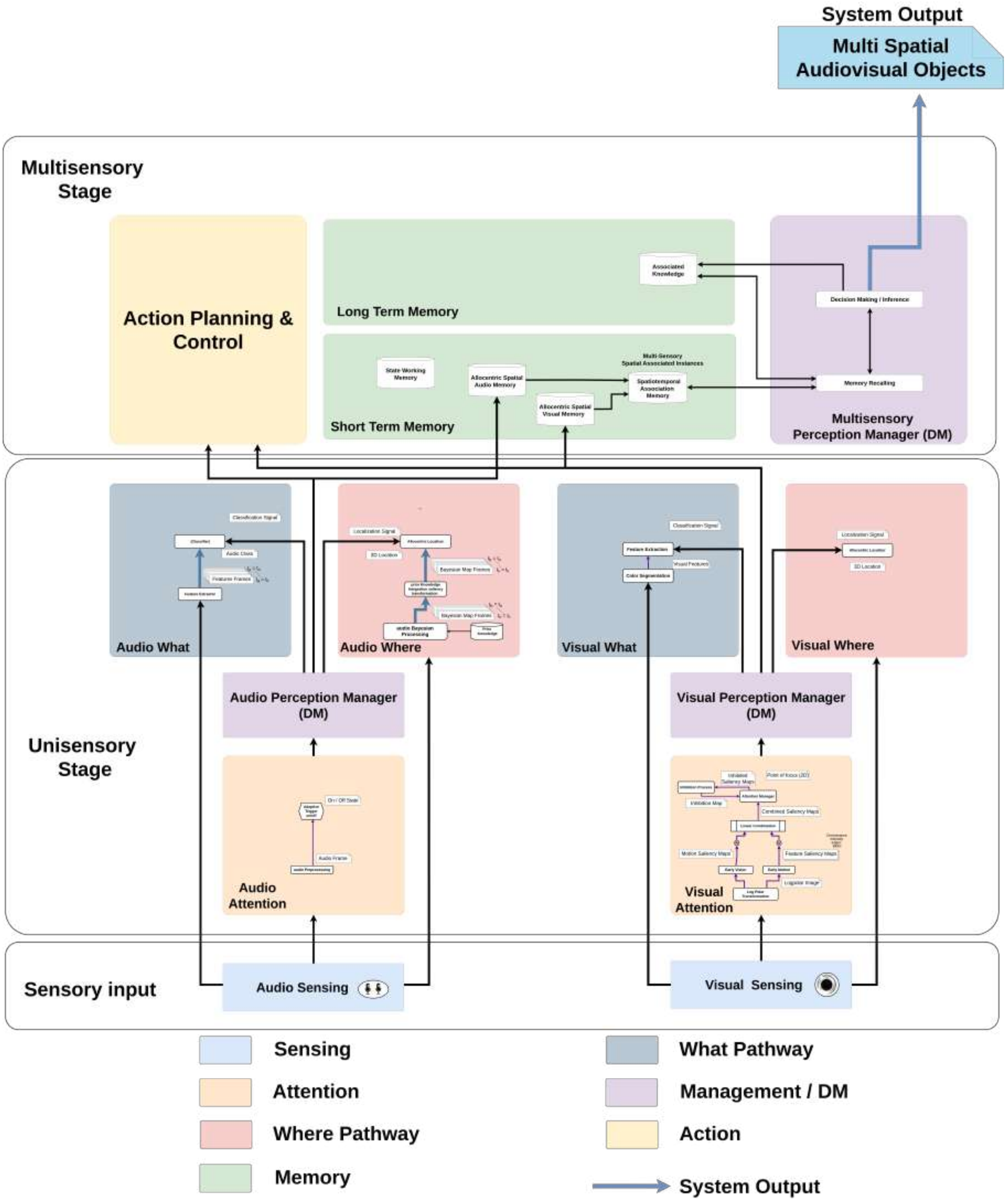


Figure 5.2 The Detailed Structure of The General Audiovisual Perception Architecture

mechanism that gives lower priority to the attended stimulus which actually helps to scan the scene and avoid keeping the focus of attention on a target (in the case this target has the highest saliency). It was first discovered by Posner and Cohen's in the 80's [189, 190]. Following this, IOR has been highly investigated and studied from different research lines in neuroscience, and cognitive behavior [142, 115]. Recently, the process of the IOR was implemented in robotics models [212, 178, 241].

The inhibition process module creates an inhibition map for each time frame. The inhibition map is integrated with the combined saliency map in the attention manager. The attention manager then gives the output which is the inhibited combined saliency map. The inhibited map is then sent to the inhibition process to close the loop between the attention manager and the inhibition process. The attention selection is then applied (by the attention manager) on the inhibited map. The output of the attention selection is a 2D point, which is the selected **hot Point** as previously presented in the first part of the thesis.

Audio Attention The audio attention in this architecture is part of the Bayesian localization model with some modifications as well. So the first module of the Bayesian localization model is the audio preprocessor. The first stage of audio preprocessing is filtering the audio to a specific range of frequencies. In this architecture, we represent auditory attention as a restriction mechanism where only a process

Based on a specifically defined range of frequencies selected and other ranges are suppressed. Following this, an adaptive power trigger is used. In the previous part, the audio power trigger was presented and used. However, the threshold, in that case, was fixed. The problem with the fixed thresholds is that they aren't adaptive to different contexts and have to be tuned to specific conditions. We addressed this drawback by implementing an adaptive threshold. The adaptive threshold is automatically tuned based on the ambient power level which is a bottom-up approach. Additionally, it accepts top-down commands from the manager (In case the agent requires a higher level of audio sensitivity). For the bottom-up process, the threshold is set based on the average and the standard deviation of the audio power over a defined period of time. The adaptive trigger constantly computes the average and the standard deviation of the power across this defined audio frame and changes the threshold accordingly. The adaptive threshold has two different modes of operation. The first mode is setting the threshold to a slightly larger value than the maximum value and the second a larger value than the average (defined by a buffer value). The update of the threshold is only made when the standard deviation is low. Using this method, the threshold is adapting the auditory seen and sets the value based on the level of the sound power when there is no stimulation.

Visual What Pathways For the visual modality, What pathway in our architecture is a simple color segmentation module and a feature extractor. The color segmentation is streaming the segments. The segments are represented with the ID of the color, and the boundary box which is defined by 2 points (top left and bottom right point). The feature extractor works as a communication channel between the manager and the color segmentation module. It receives a request command with a 2D point (Point of focus) and then replies with the features of this point/region. It replies with the color and the area of the focused point as well as a confidence value. The confidence value is calculated based on the temporal changes at this point in the scene. It is accumulated over time frames. So for example, in the case of observations, while the robot is moving, the scene will be changing over time. Consequently, the confidence rate will be low. To summarize this part, the visual what pathway is analyzing the features of the scene and replying to the request from the managers with a set of features and a confidence level for these features. This confidence level is computed and updated temporally.

Auditory What Pathways On the other hand, the auditory what pathway is a features extractor and a pre-trained audio classification module. The classification module receives a classification request signal from the audio manager. Once received, the classifier replies with the predicted class of the audio with a confidence rate. The classifier is also receiving a stream of features for an audio signal with a buffer size. As the auditory signal requires some time frames to be processed (unlike the visual image), the classifier doesn't reply immediately to the manager. Instead, it takes some time (1 second minimum) to observe the sound and try to recognize it. Further, it keeps classifying the auditory signal until the manager sends a stop signal. The classifier used in our implementation was a trained SVM model to classify the four kinds of complex wave signals. It also gives a confidence level.

Visual Where Pathways The visual where pathway is doing an allocentric localization in the 3D space. The output is a 3D position in the space for the 2D point of focus. For simplicity, we used a 3D projection method which was adapted before in the previous version of the architecture in part II of the thesis.

Auditory Where Pathways Similarly to the visual where pathway, we used the Bayesian localization model with the integration of the prior knowledge that was previously implemented and used in part II of the thesis. However, in this version, another module was added which is the audio allocentric localizer. This module localization is using the time-variant decision-making concept that was implemented before to localize the auditory stimulation. This is possible with using the use of working memory.

Visual and Auditory Managers The processes of the visual and audio managers are identical. They perceive a signal from the attention system. Afterward, they acquire data from both what pathway and where pathway by sending a request to each of the pathways. After sending signals to the pathways they wait for the replies from the pathways. Once received a reply from what and where they bind the data together into an object and send it to the short-term memory. The object consists of three main elements, the 3D spatial location, the features set, and finally the confidences. There are three confidences associated with the object. The first confidence is spatial confidence which represents how accurately the robot was localizing this stimulus. The second confidence is class confidence which represents how accurate the robot is in classifying this stimulus. And finally, the object confidence which represents the overall confidence about this object based on equation 5.1. W-what and W-where are the weights of the importance of the what pathway and where pathway. They are parameters for the architecture. We used equal weights (0.5 for each of them) in our experimental settings.

$$Con_{obj} = W_{what} * Con_{what} + W_{where} * Con_{where} \tag{5.1}$$

Confidence of a Unisensory Object Instance (Visual/Audio Manager)

5.2.2 Short-term Memory

Short-term memory holds 4 kinds of memory. The state working memory, the allocentric spatial visual memory, the allocentric spatial auditory memory, and finally the spatiotemporal association memory.

State Working Memory The state working memory is the one implemented and used in the previous versions of the architecture in part II of the thesis. It supports the decision-making process in localization and tracks the state of the stimulation and the action. It also tracks the state of the multisensory-associated object. The state is defined by the existence of the stimulation and automatically set by false after a defined period.

Allocentric Spatial Visual/Auditory Memory The allocentric spatial memory is the place where unassociated objects are preserved for a short time. The unisensory object is perceived from the manager module, and then the object is added to the memory. The adding process is done through a Bayesian integration. In this process, the priors are integrated with the received stimuli in a Bayesian integration if they have the same class and are perceived within a defined range of distance in space. These are the two conditions that have to be met to perform the Bayesian integration process. Otherwise, the perceived object is considered another new object in space. The Bayesian integration is done based on the confidence of the priors and the current confidence using the equation 5.2. The Pos can be one of any of the axis (X, Y, Z) in the cartesian space in the case of Vision, and angels and radius ($\aleph$, β, R)in the spherical representation. Further, the posterior confidence is then calculated. Also the confidence is updated using equation 5.3. It is a weighted summation of the confidence noting that it is normalized. This means that the summation of the weights should be equal to one.

$$Pos_{new} = \frac{Con_{prior}}{Con_{prior} + Con_{current}} * Pos_{prior} + \frac{Con_{current}}{Con_{prior} + Con_{current}} * Pos_{current} \tag{5.2}$$

Spatial Position Equation For The Unimodal Instance Considering The Priors (Visual/Audio Spatiotemporal Memory)

$$Con_{new} = W_{prior} * Con_{prior} + W_{current} * Con_{current} \tag{5.3}$$

Prior Experience Inference For The Unimodal Instance Considering The Priors (Visual/Audio Spatiotemporal Memory)

Spatiotemporal Association Memory In the spatiotemporal associated memory, the objects from the sensories are integrated with all possibilities. The associations are created, between each visual object with each auditory object. The association is done based on equation 5.4 and the final location of the associated object is computed based on equation 5.5.

The associated instances here are representing all possible combinations of the existing unisensory instances and the memory doesn't take a perceptual decision on the final association. It is just an initial stage of association. The perceptual decision is taken by the manager which will be addressed in subsection 5.2.4

$$Con_{association} = W_{visual} * Con_{visual} + W_{audio} * Con_{audio} + W_{distance} * d(Obj_{audio}, Obj_{vision}) \tag{5.4}$$

Associated Confidence Equation For the Multisensory Associated Instances (Short-Term Memory)

$$Pos_{associated} = \frac{Con_{vision}}{Con_{vision} + Con_{audio}} * Pos_{vision} + \frac{Con_{audio}}{Con_{vision} + Con_{audio}} * Pos_{audio} \tag{5.5}$$

Spatial Position Equation For the Multisensory Associated Instances (Short-Term Memory)

The spatiotemporal association memory also matches the visual space (which is represented in Cartesian space) with the auditory space (which is represented in spherical space). Thanks to the allocentric localizers of both modalities, both coordinates share the same origin. Technically, this is done by knowing the internal state of the robot's head angles (including the eyes' positions).

One important point related to short-term memory is the elimination of the content. We implemented this based on time constraints. Each instance will be removed if it remains in the memory for a defined amount of time.

5.2.3 Long-term Memory

Long-term memory stores the past binding states of the multisensory features of the perceived objects. This information is stored as a list of items. Each Item has its visual features, auditory features, accumulated confidence, and count of the perceived experiences of this association. The perceived instances are decided by the multisensory manager module. The manager sends the experience to long-term memory. The experience has visual features, audio features, and confidence. If the features are matched with a previous experience, the count is incremented by one and the confidence is updated according to the equation 5.6. This is implemented in the Associated Experiences memory module.

$$\frac{Count_{old} * Con_{old} + Con_{new}}{Count_{old} + 1} \tag{5.6}$$

Accumulated Confidence Update Equation (Long-Term Memory)

5.2.4 Multisensory Manager / Decision Making

The multisensory manager is a connection between short-term memory, where the instances are stored temporarily, and long-term memory where the associated experiences are stored. The manager has two modules. The memory recalling module, and the multisensory decision-making/inference module.

Memory Recalling The memory recalling module deals with both short-term and long-term memory. It recalls the whole content of the spatiotemporal association memory in each frame. Also, it sends a list of queries to the long-term memory component (the associated Experience Module) to recall the experiences that have the same associated features (for each possible association that was recalled from the short-term memory). The query has both the visual features and the audio features of the associated instance. The reply from long-term memory is the accumulated confidence and count this associated instance was perceived in the past. If the experience is completely new, it will receive zero for both the confidence and the count. Further, the set of possible associations (from the spatial association memory) and the prior experiences (from the associated long-term experiences) are sent to the decision-making module.

The Decision-Making The decision-making module applies prior inference to each of the possible associations based on equation 5.7. This process ranks the associated possibilities. This rank is then used in the decision. The objects are decided based on this rank. After deciding D on an association that has a visual object V and an audio object A, the associations that consider the unisensory objects V and A in the association list are dropped as they have already been considered. The count of the possibilities is then decreased to (N-1)*(M-1). This process of selection and dropping until the list of association probabilities is finished. In case of unequal count for the visual and auditory count, there will be remaining unassociated objects. These objects are considered in the perceptual decision as a unisensory experience.

$$Rank = W_{prior} * Con_{experience} + W_{current} * Con_{instance} + W_{count} * Count \tag{5.7}$$

Prior Experience Inference Ranking Equation

Conscious Output The output of the architecture is what I called conscious output. It is the output of the spatiotemporal integration and inference processing of the prior knowledge that includes the statistical characteristics of perception and previously learned experiences. It is a representation of what the robot is perceiving at each time frame. It is a collection of objects and their spatial location in space.

5.2.5 Incremental Approach

Overall, the architecture goes from the low level to higher level processing by passing through an attention stage for each of the unisensory modalities (Vision and Audio), then two parallel processes as dual pathways (what and where) for each modality. Further, the manager controllers the dual pathways and the attention components, acquire the spatial location from the where pathway, acquires the features from the what pathway, temporally integrates the acquired data (reply from what pathway and the reply from where pathway), and finally creates an object. Both managers (Audio manager, and visual manager) are then sending the object to the spatiotemporal short-term memory. The memory temporarily holds the required information about the objects with their perceived confidence. The spatial position and the confidence are also updated in the memory following a Bayesian way (multiple instances from the managers are perceived). In the spatial association memory, each existing object in the unisensory spatiotemporal memory is associated with all objects from the other modality as association possibilities. This is generating N*M possibility for the associations where N is the count of the visual instances and M is the count of the auditory instances. Each associated possibility has a degree of confidence based on spatial factors and the confidence of the object from each modality. The closer the perceived location the higher the confidence value of the association in the associated memory. At this point, the decision on the association is not yet finalized. It only considers the structural factors that come from the environment. The multisensory manager block then acquires the associated set of multisensory objects with their confidences. Additionally, it requires the learned associations (which were stored from previous decisions) from long-term memory. The priors from the long-term memory bias the confidence of the associated possible sets and change their confidence values. Further, the conscious decision on the whole scene is made based on the biased confidence values. Once an associated instance id been decided, some other instances are dropped (because the unisensory object is already associated). Finally, the manager updates the long-term associated experience with the final decision to update the stored experiences. The working memory component tracks the states of the stimulation and the robot which regulate the

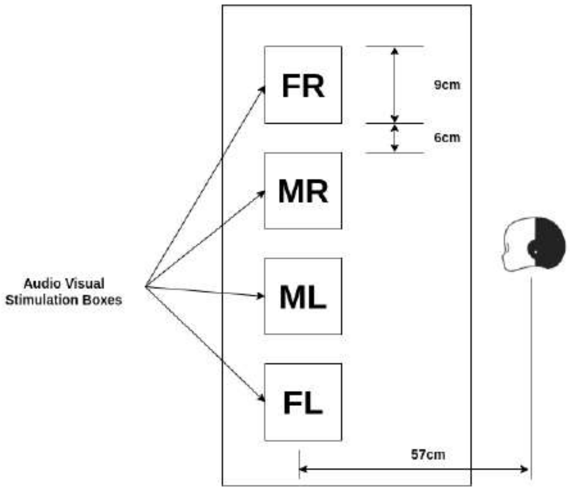


Figure 5.3 The Experimental Setup To Evaluate the Performance of The Proposed Cognitive Architecture For Perception

decision-making process of the unisensory managers as well as the multisensory manager. This facilitates the update of the memory only once for each stimulation event.

5.3 The Experiment

This experiment aims to examine the architecture as a whole and see if the robot will be capable of understanding the presented scene using this architecture. We designed this cognitive architecture to enable cross-modal interaction. In particular the interaction in the localization and identification task. The specific point that we would like to examine in this experiment is whether if the architecture is capable of capturing the association of the presented features of the stimulus in the scene and using these features to solve an ambiguity in the scene.

5.3.1 The Setup

The setup that was designed to test this architecture was focusing on capturing the right stimulus on its location. The setup used in the experiment of chapter 4 was used but without human participants, as shown in Figure 5.3. In particular, four audiovisual stimulation boxes were placed in front of the robot (all in a single line, with a 6cm difference between the boxes which is 15 cm center to center). The experiment consists of two phases that will be explained in section 5.3.3.

Table 5.1 The audiovisual stimulation objects with their IDs and features

Object ID	Vision		Auditory	
	Visual Feature	**ID**	**Audio Feature**	**ID**
0	Red	0	1.5 kHz with 3 harmonics	0
1	Green	1	1 kHz with 3 harmonics	1
2	Yellow	2	500 Hz with 3 harmonics	2
3	Blue	3	250 Hz with 3 harmonics	3

5.3.2 Stimulation

The main focus of the experiment is to examine the accuracy of the system in localization and classification and how the association of the features will be learned by the system. In this experiment, we adopted four different audio signals and four different visual colors. Additionally, we associated each color with an audio signal to result in four different audiovisual stimulation as shown in Table 5.1.

5.3.3 The Experimental Phases

As mentioned the experiment is divided into two phases, the first phase is to learn the association, and the second phase is for testing what the robot learns and whether the architecture would be able to solve a complex scene. We recorded the final output of the robot (The conscious output). Additionally, the long-term memory and its updates across time were also recorded. Indeed, this is the final output of the system, and many other inside nodes (modules) could be recorded and tested. However, we focused only on the final output of the whole architecture to simplify the analysis. In the future, we aim to dig deeper in the analysis of each single module and show its behavior. The main task of the robot is perceiving the scene and describing which objects and cues are presented and where in space. As part of the architecture, detecting the features of the objects (Colors for vision, Complex tone in audio) is one of the prior capabilities as the robot uses trained modules in the architecture. However, the robot has to learn which cues are associated together to create an internal representation of the object throughout the stages of the experiment. The multmodal objects are shown in Table 5.1. Additionally, we visualize the process of both phases in Figure 5.4 and will explain the phases as the following:

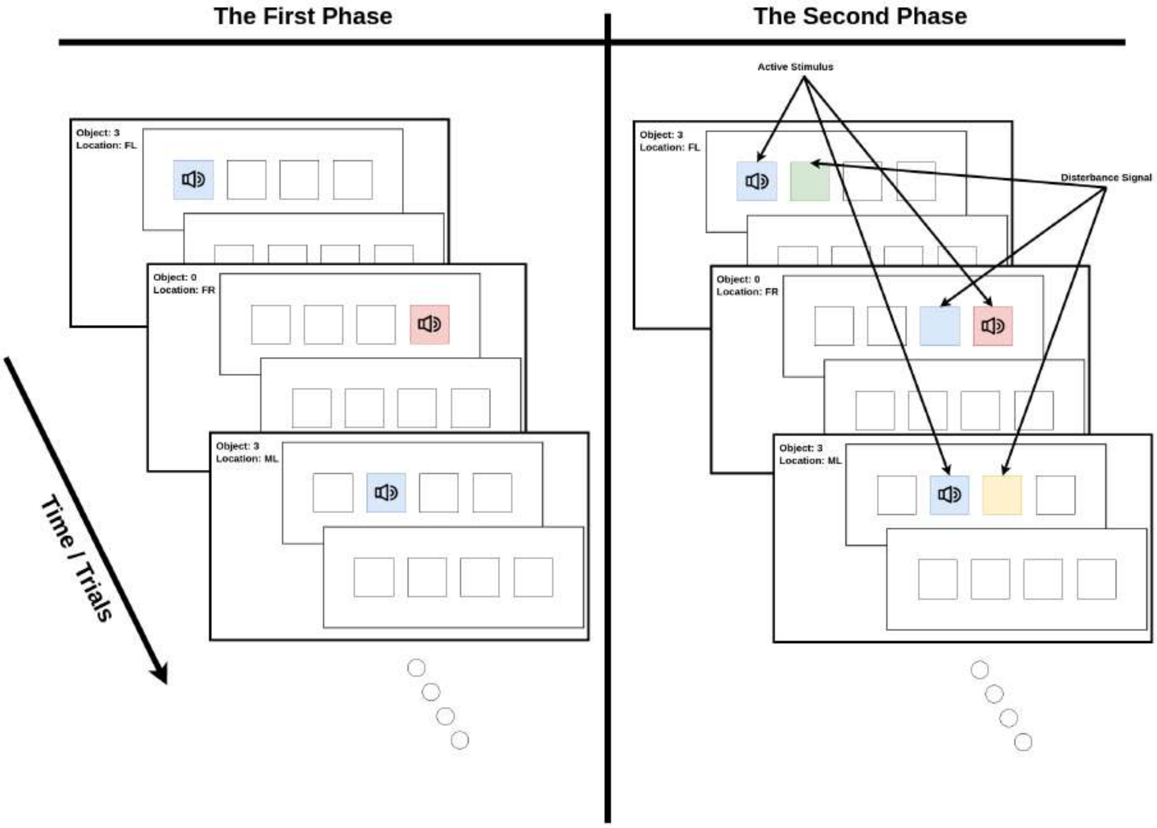


Figure 5.4 The Stimulus Behaviour During The Two Phases of The Experiment

Phase one: Learning the association The first phase of the experiment was turning on one object for both modalities at one of the four locations. The selection of the object and the place is random. This phase of the experiment consists of 100 trials. The trial is 10 seconds of on-set time and 10 seconds of off-set time. The aim of this phase for the robot is to create an internal representation of the associated features as a single multimodal object with a confidence level. The difficulty of this task is to correctly localize the audio stimulation and the visual stimulation and integrate them together in a single object. Further, use the knowledge acquired from the experiences to build up a confidence level starting from a low level of confidence as firstly the robot doesn't have any, to an increased level of confidence when the experience is repeated.

Phase two: Cross-binding The second phase of the experiment is more challenging for the robot. One audiovisual stimulus will be presented but this time with the existence of a visual disturbance signal. The disturbance signal is a visual-only stimulus that is always from one of the neighbor's stimulation boxes. (the right or left box next to the audio-visual stimulation).

The selection of the position (The active box/s) was random. The only constraint was that the audiovisual stimulation and its disturbance has to be next to each other with no box in between. The main aim of this phase is to examine if the architecture would be capable of capturing the right audiovisual stimulation and understanding the scene correctly. This task is a complex one as it requires solving a cross-binding problem. The robot has to decide which visual object is associated with the auditory signal. It is also based on the learned associations in the first phase. The efficiency of the first phase will directly affect the performance of the robot in the second phase.

5.4 Results & Discussion

The results of the experiment are divided into two components. The first component is from the first phase of the experiment which is the internal representation of the presented objects. On the other hand, the second component is from the second phase which is the accuracy of the output of the model in the cross-binding scenarios.

5.4.1 Object Internal Representation (First Phase)

Figure 5.5 shows the accumulated confidence profile. The confidence profile is the different confidence levels of the objects in long-term memory across the trials of the experiment. The figure shows seven lines which means that robot learns seven objects indexed from 0 to 6. Each of the lines is representing a multisensory object which is a bundle of audio features and visual features. In our case, the visual feature is the color, and the audio feature is the sound class as previously explained. Three of seven objects (in the graph) had a low unchanged value. Also, they appeared at different points in time (trial). The starting point of the line represents the trial in which this object was perceived for the first time. The values of the other four objects were increasing until they reach confidence equal to one. which is a saturated value for confidence. This means that the robot learned a strong association between visual features and audio features for these four objects. To understand more what are these features and whether the learned associations are correct, I add here Table 5.2 which shows the latest version of the long-term memory. Comparing the values in the table with the correct values shown previously in Table 5.1, we can identify the wrong associations which are marked in gray in the Table. Also, we can observe the strongly evident associations (the four objects that reached the saturation level) are the correct objects. The wrong associations happened due to a false classification for the audio class or the visual class. These results

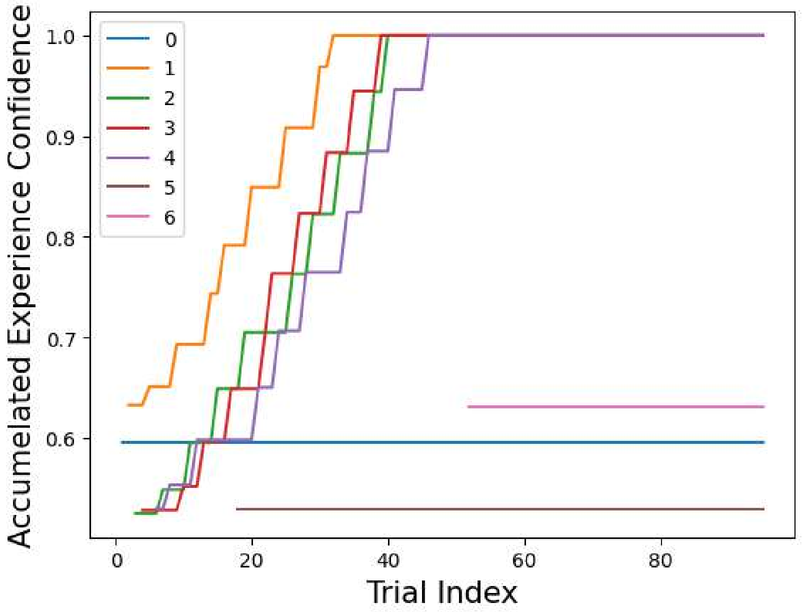


Figure 5.5 The Confidence Profile For The Objects in The Long-term Memory

demonstrate that the architecture endowed the robot with the capability of associating the correct features together and learning the multisensory perceived objects. This learning process is dependent on multiple internal processes in the architecture including the dual pathways of both audio and visual modalities, the spatial working memory, the perceptual inference, and the prior integration from the long-term memory.

Table 5.2 The Saved Object Features Associations in The Long-term Memory At The End of The First Phase of The Experiment. The Wrong Associations Are Marked in gray background.

id	Visual Feature	Audio Feature	Experience Count	Accumelated Confidence
0	BLUE	2	1	0.596003
1	RED	0	23	1.000000
2	GREEN	1	23	1.000000
3	BLUE	3	24	1.000000
4	YELLOW	2	21	1.000000
5	YELLOW	3	1	0.529363
6	GREEN	2	1	0.630290

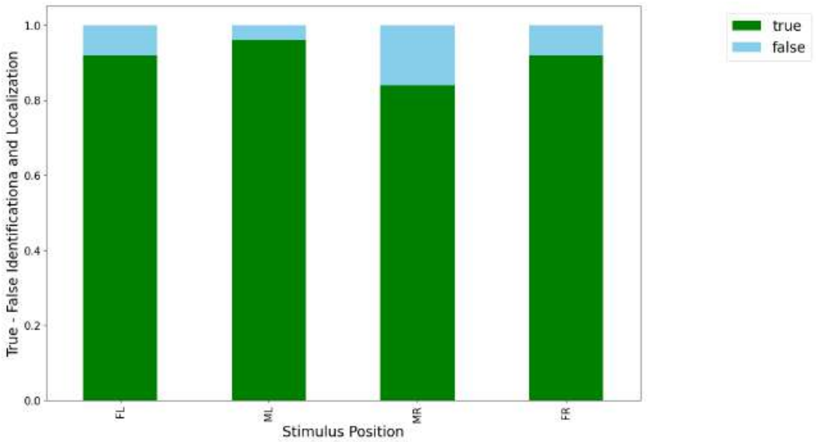


Figure 5.6 The Accuracy of the Identification and Localization of the Stimulus Across Their Different Locations

5.4.2 Accuracy of Solving the Binding Problem (Second Phase)

In the second phase, the presented scenes to the robot were more complex compared to the first phase. In the second phase, the robot has to identify and localize the audio-visual object as well as the disturbance (a visual object next to the audiovisual object). We analyzed the output of the system which is the final decision of the robot for the perceived environment. The output is a list of objects and their corresponding spatial location. Each object can be an audio object or visual object or audio-visual object. Figure 5.6 is showing the accuracy of the scene description of the robot. The description is considered true if the robot recognizes and localizes both the stimulation and the disturbance in the right way. The stimulation is right if the output includes the object with the correct audio-visual features and in the correct spatial position. And the disturbance is correct if it is recognized and localized correctly and not associated with an auditory feature. This is the correct binding way in which the robot binds the auditory feature with the correct visual feature in the scene. For localization, it is considered right if the error is below 5 cm (Outside the frame of the stimulation box) in each plan (X, Y, Z). Combining all stimulus locations (FR, MR, ML, FL), the architecture achieved 91% accuracy in the second phase. The figure is showing the accuracies for each location which are all above 83% accuracy.

The experiment aimed to examine the overall behavior of the architecture which is our main focus. In particular to evaluate if this architecture would give the robot the capability of resolving a complex scene in which the robot has to integrate or segregate the features from different modalities (auditory and vision). The analyzed results of the experiment that the robot was capable of resolving the complex presented scenes and describing them correctly. The proposed architecture is biologically inspired and implemented multiple biological processes internally. The experiment showed the efficiency of this approach. This architecture is a dynamic one that is not trivial to evaluate every single module such as the Inhibition of Return (IOR). More specified experiments can be performed on this architecture to show the behavior of every single module within this architecture in real-time on the robot and also in simulation.

5.5 Conclusion

One of the important skills for the robot to share the perception with a human partner is to have a common representation of the environment. This means that the robot creates a similar understanding of the scene with respect to human understanding. This understanding goes beyond object recognition and auditory classification and includes the spatiotemporal dynamics of the processing mechanism. Therefore, we proposed a cognitive architecture that does audiovisual processing for the scene inspired by the biological process. The process starts from the previously developed architecture in Part II with some modifications and added processes. In brief, the architecture starts with attention processing for each modality. Following this, dual unimodal pathways (What, and where) are processing the data under the organization process and decision-making that are processed by the manager. The manager is then integrating the information from the pathways into an object taking into consideration the confidence processing. The objects are then sent to the spatiotemporal working memory, where Bayesian processes are applied to create a set of possible associations. Further, the multisensory manager decides about the scene and creates an output which is what we called the conscious output. This output is considering the long-term associations that are stored in the long-term memory as priors and consider them in the decision-making process. The output (conscious output) of the architecture is a set of objects that are present in the scene which is given for each time frame.

The architecture was examined in two phases. The first phase was to examine two behaviors. The first behavior is understanding a simple environment (One audiovisual object) and the second behavior is learning the associations. On the other hand, the second phase was focused on examing the behavior of the architecture while presenting a complex cross-binding scene (One audiovisual object + visual disturbance). The result of the first phase showed a learning curve for the association until it reached saturation for the right features. In the second phase, the robot was capable of solving the complex scenario and giving the right expected output which accurately describes the presented objects. The accuracy of the system across both phases was 91% considering both the localization and recognition processes. The proposed architecture can be technically applied to any perceptual audio-visual process. It is not specified for a specific task such as human localization or navigation.

My contribution to this chapter is in the design and implementation of the audio-visual perception architecture. The main novelty of this architecture is in the following four points:

- Consideration of the cross-modal interaction in perception cognitive processing
- Integrating the prior knowledge and experiences in different levels of perception (auditory localization processing, spatiotemporal dynamics in the short-term unimodal processing, and in association using the long-term memory).
- The time-variant and confidence-based decision-making process in the managers (visual perception manager, auditory perception manager, and multisensory perception manager).
- The generalization of our approach. It is a general modal, that is not limited to a specified application. From one application to another, the architecture can be used by only tuning the attention parameters and embedding the required recognition modules for both audio and vision. (An applied example will be explained in chapter 7.

The limitation of this architecture is only in the auditory localization processing. Generally, it takes around 5 seconds to converge. However, with this slow response, the robot was able to rely on the other processing pathways to perceive the environment correctly. This shows the strength of this approach.

The experiment explained in this chapter shows the overall behavior of the architecture. However, there are much more experiments that can be performed in the future to show the internal processing of each module and the internal dynamics. It is important to note that the architecture is very modular which is another strength of this architecture. It can be used in different applications to choose the required process and disable the pathways that are not useful in the designed task. This will be demonstrated in chapter 7 which is the work done to answer the third research question *RQ3*.

Using this architecture the robot was able to learn the association over time and used this association to create a conscious understanding of the scene and handle the uncertainty. Further, more complex scenarios can be examined using this architecture. such as using a higher level of feature space, or examining the illusions such as ventriloquist as a benchmark experimental behavior vs human behavior. The presented work in this chapter answered the first two parts of the second research question *RQ2*. The last part of the research question is about how the robot can actively perceive the environment. The architecture and the implementation are considering the active component as shown in figure 5.1, there are action planning and execution component. However, the next chapter will explain our research work in this direction. The research work explained in the following chapter is focused on proactive behavior specifically for improving auditory localization skills as it is the main limitation of this architecture.

Chapter 6

Enabling Proactive Strategies in The Architecture to Improve Perception

> I hear and I forget. I see and I remember. I do and I understand.
>
> **Confucius**

In the previous chapter, we proposed a cognitive architecture for perception. The action component in the proposed architecture was already represented. However, the experiment didn't explore the proactive role of the architecture. The proactive role of the architecture is the action that is initiated internally to improve the performance of the perception. In this chapter, we focus on this role. Specifically, for auditory localization. In this research part, we try to address human-like proactive behavior in the perceptual process. The research presented in this chapter is to tackle a part of the second research question which is *How can the robot actively perceive the environment?*. The active part of perception is engaging the action in the process of perception. Looking at the shared perception skill categories, this work is targeting a common proactive behavior in the perception process with humans.

6.1 Introduction

In typical human-human interactions, most of our sensorial inputs coregister to provide a complete understanding of the multisensory context in which we interact. In human-robot interaction, visual perception has been widely addressed, whereas other perceptual channels such as auditory perception are less well-studied. The auditory scene is dense with information about the people and their activities around a robot. This is true even for sound sources that are inaccessible to vision, such as around corners and off the field of view of the vision sensors. To make optimal use of this information, the spatial locations of auditory events must be correctly perceived and registered into an allocentric reference frame around the robot. Substantial research on this localization problem has yielded improving results for both binaural and microphone array systems [8] [197]. An emerging theme in this research is the need to exploit the complex interaction between the robot's movements over time and the continuously updating memory of sound locations in the space around the robot. Binaural auditory systems, both biological and artificial, perform auditory scene analysis by computing either (or both) interaural time differences (ITD) and interaural level differences (ILD). However, such systems necessarily produce "phantom" images of sounds due to ambiguities in ITD and ILD computation [21]. One solution used by biological systems was proposed by Hans Wallach [249] [250], which involves integrating information across head rotations. This active-hearing approach has been successfully used (e.g. by [79],[17], [259], and [124]) to better resolve the sources in the scene. The process typically engages head rotations and the integration of beliefs during the rotation. Optimal head movements thus not only reorient the hearing system to better perceive the sound sources but also improve auditory localization [79]. A key aspect of active-hearing approaches is that they use some variation on Bayesian memory (e.g. via Kalman-like or particle filters) and cannot exclusively rely on instantaneous evidence. Information about the pose of the robot, as it changes in allocentric space over time, must be integrated with egocentric instantaneous evidence about the auditory scene, to produce a posterior probability map of auditory objects. In fact, the prior expectation of auditory information in the auditory scene is a crucial element in human hearing, especially for priors' expectations extracted from vision. For example, the integration of visual evidence with auditory perception gives rise to the well-known ventriloquist illusion of sound localization [4] and the McGurk Effect in audiovisual speech perception [148]. In the literature, the style of motor actions that leads to optimal performance has not been studied yet.

In this chapter, we propose a cognitive framework that includes motor control and working memory modules to improve auditory localization in a localization task. This architecture is part of our proposed cognitive architecture that was proposed in the previous chapter (chapter 5). It uses auditory attention, auditory where pathway, the state working memory, and action execution. The aim of this sub-architecture is to present the proactive component of our proposed architecture in one of the perceptual tasks and show the modularity of our proposed architecture. Additionally, we aim to examine the possibility of using this architecture to examine the possibility of studying motor strategies that improve sound source localization

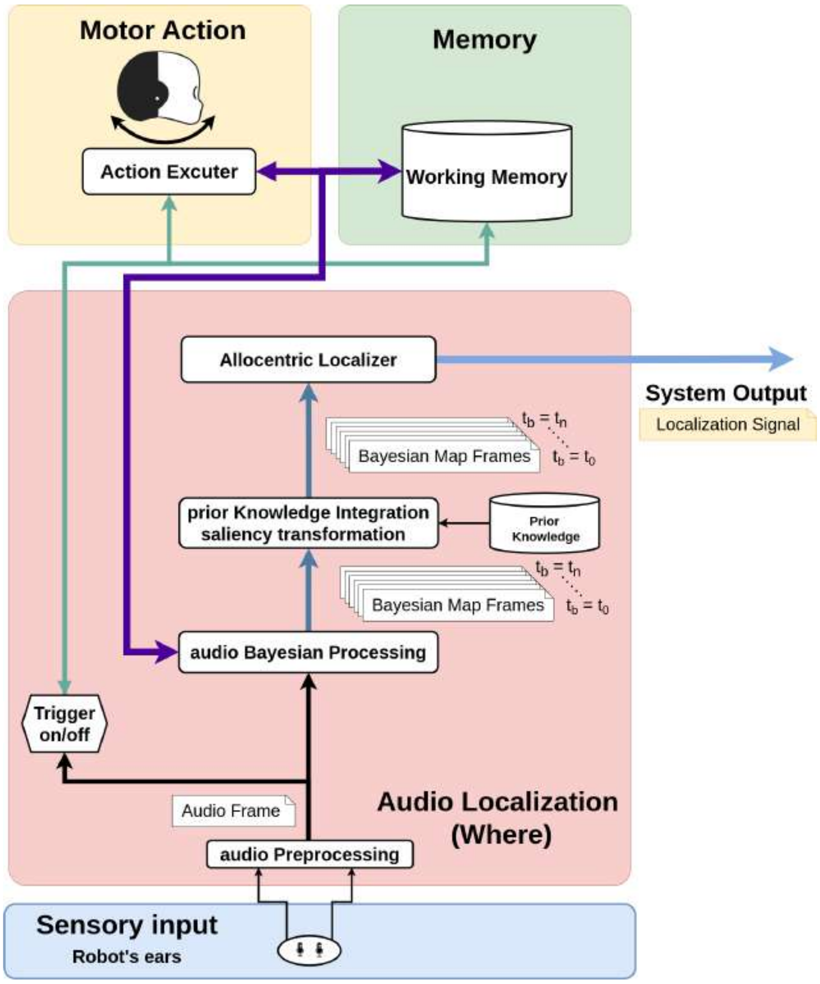


Figure 6.1 System diagram

(the effect of different head motor actions on the performance of sound localization). Here we report preliminary evidence that will inform future choices of motor behavior models.

6.2 The Sub-Architecture

Figure 6.1 shows the structure of our proposed framework (Sub-Architecture). It consists of four main elements: 1) *Sensory input*: the two microphones located on the head of the robot; 2) *Memory*: the state working memory element; 3) *Audio localization*: the auditory where component; and 4) *Motor actions*. Nothing particularly changed in the modules of the architecture compared to the version in the previous chapter (chapter 5. Briefly, the approach uses a gammatone filter bank to spectrally decompose a sound into narrow frequency bands, and a series of swept narrow-band beamformers that approximate the

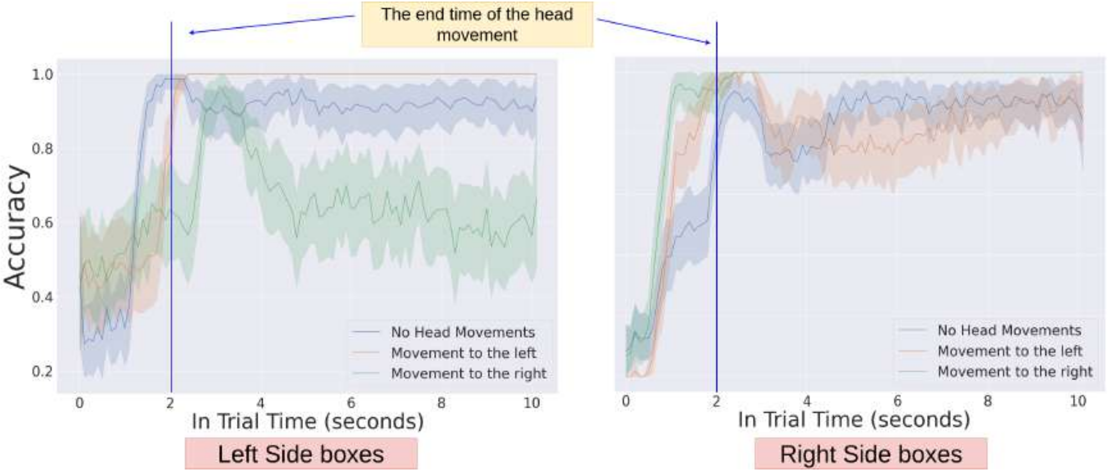


Figure 6.2 Average accuracy & std in the three motion conditions. Blue lines indicate the end time of the movement.

binaural temporal comparisons of ITD. Instantaneous egocentric auditory scenes are rotated in (*audio preprocessing*) module and used to update an allocentric Bayesian posterior map of probable sound sources in the 360 degrees around the robot in (*audio Bayesian processing*) module. The integration of a priori information about the auditory scene is processed in (*prior knowledge integration*) module. This module is biasing the system output toward known possible target locations. The system also solves the problem of sound onset detection, by computing the total power of the input sound and identifying the existence of a sound signal if the total power exceeds a predefined threshold. The off trigger is activated when the power is not exceeding the threshold. The working memory saves the priori information, the action state (Not executed/In progress/executed), and the state of the sound (Not present/ present). The trigger sets and resets the state of the sound in the working memory. This trigger is connected also to the Action Linker, which executes a pre-defined motor policy based on the trigger and the state of the action from the working memory.

6.3 Preliminary Experiment and Results

On each trial, a complex tone (1KHz with 3 harmonics) was presented for 10 seconds from one of four identical boxes located horizontally on a table in front of the robot (-23° , -8°, 8°, 23° with respect to the midline). The robot's task was to determine which of these boxes produced the target sound. Three movement conditions were considered: no head movement, rotation of the midline toward the direction of the target, and rotation away from the target. Rotation was at 5 degrees/seconds for 2 seconds. The robot completed 24 trials for sounds from each box. From the allocentric posterior map described above, we extracted the azimuth of maximum probability as it changed over time and expressed this as percentage of time points at which the max probability corresponded to the target sound location. The results in Fig. 6.2 show that the accuracy depends on both time and direction of head rotation: In all conditions, accuracy improves quickly from chance but stabilized at 100% only when the head rotated toward the target. Other aspects of the system were stable. The system detected the sound almost instantaneously and the motor actions were performed correctly.

6.4 Discussion

We proposed a framework to empower the robot with motor actions with the target of enhancing the audio perception of the robot. Demonstrating our concept practically, for example, the model can be applied in industrial applications that use an active auditory signal to promote interaction between the industrial robot and the human operator. Our work will directly improve the interaction with the human by better accurate audio localization for the sound signal in the working environment as the human localize. We tested our implementation on the iCub robot and explored the effect of specific head movements: the azimuthal rotation of the robot head (yaw) in two directions, relative to static pose. Overall the frameworks worked well. The audio power-based trigger successfully captured the starting of the sound signal, both signal and action state were tracked in the working memory component correctly, and both the action and reset to the home position were executed in the correct timing. Regarding the performance, the experiment showed encouraging results regarding the use of prior information and motor actions: the system converged from chance accuracy (25%) to near-perfect accuracy (100%) within about 2 seconds, particularly for the optimal motion strategy. Interestingly, during the initial seconds of the trial, the solution tended to be less stable, and the static pose also improved considerably from chance. This means that time - and not head movements - is sufficient for a Bayesian system to achieve relatively good localization in this task. However, beyond about two seconds, head rotation toward, but not away from the target was necessary to achieve stable 100% accuracy. Head rotations away from the target were not as useful, suggesting that other factors such as the speed of the motor action and the initial orientation of the head before sound onset might also be consequential.

This study provides insight into how auditory Artificial Intelligence might provide relevant details about the world and human partners in the HRI context. Given that complex human-robot interactions might present conflicting requirements for the robot to orient its microphones both toward a target (for optimal localization) and away from a target (in response to an instruction or to accomplish some other behavior) it will be interesting to explore the integration of various motor behaviors into active auditory perception. More importantly, the evidence that motor commands improve auditory perception in humans gives us confidence that exploiting proactive behavior for robots will help improve the next generation of perceptual skills in interactive scenarios.

6.5 Conclusion

In this chapter, we proposed an architecture to enable proactive behavior for the robot in auditory localization. We choose auditory localization as it is the most challenging task in the proposed architecture compared to other tasks like visual localization or auditory classification (as previously shown in previous chapters). The proposed architecture in this chapter is a partial part of the whole model proposed in chapter the previous chapter. We showed how can proactive actions be activated and further improve the perception (audio localization in this case). The actions are inspired by human behavior of doing head movements to improve their auditory perception. This work is integrated into the main perception architecture presented in the previous chapter (chapter 5. By this, we finalize the research work for the second research question *RQ2*. The following chapter will present the main work done to address the last research question *RQ3* which is about generalization and coordination in a real-world application.

Chapter 7

From Sensing to High-level Planning and Coordination: Generalization and Action Coordination in a Real-world Application

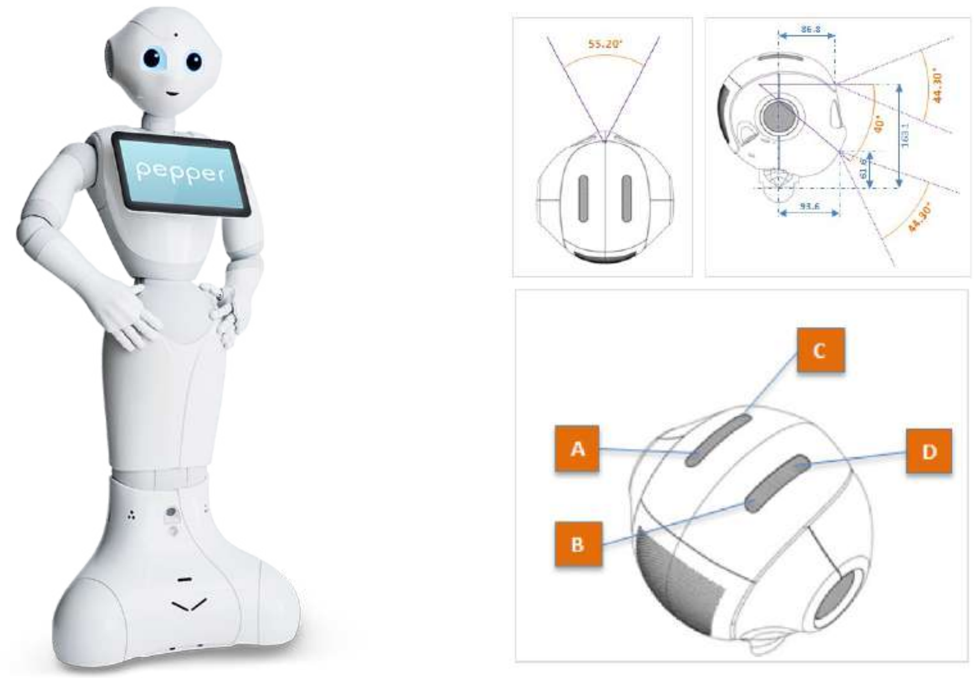
I like to think of ideas as potential energy.
They're really wonderful, but nothing will
happen until we risk putting them into action.

Mae Jemison

In the previous chapters, we proposed a cognitive architecture for active audio-visual perception. Going further one step in the research, we present in the chapter the research carried out to answer the last research question *RQ3: Can this perception architecture be generalized to different robots and applied to a complex task that requires coordination with another agent?*.

7.1 Generalization on different Robots

The aim of this work is to examine if the architecture is general enough to be used on different robotic systems that have different embodiments and different software. This part was done in the first abroad period at the University of Essex, UK. The work was done in two other robots which are Pepper, and the Essex agricultural robot. In the following part, a brief introduction about the robots will be explained followed by the implementation details to adopt the architecture of these two robots.



(a) Fill Body

(b) Microphones and 2D Cameras

Figure 7.1 Pepper Robot

Pepper Robot Pepper is a humanoid robot that was developed by SoftBank (shown in Figure 7.1a. It is equipped with many sensors and actuators. It has a depth camera (the eyes) and two HD cameras in the head designed to be used in visual tasks. One Camera is placed in the center above the eyes, and the other one is in the mouth of the robot. Both cameras are covering different fields as shown in 7.1b. For audio, Pepper has four microphones placed as a 2*2 matrix in the head as shown in figure 7.1b.

From the software perspective, Pepper has different software versions. The older version (2.5 and older) is using Naoqi SDK, which allows direct access to the sensor readings and more flexibility in controlling the robot. The software also allows the writing of customized modules using python and C++. The new software of Pepper (2.9) is using Pepper SDK (QiSDK) which is typically programmed as a Native android application and installed on the tablet placed in the chest of the robot. The new software provides higher-level functions and limits direct access to the sensors.

Beast Robot The beast robot shown in figure 7.2 is an agricultural mobile robot. It is an assembled robotic system using multiple research and industrial robots from the market. The base is a Husky robot. And it is equipped with 2 identical UR3 robotic arms. The robot is

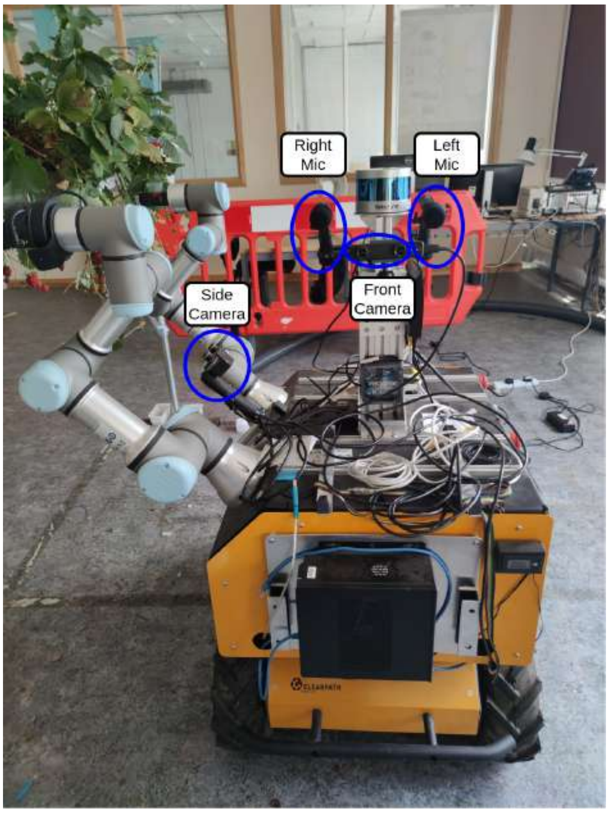


Figure 7.2 The Essex Agricultural Robot (The Beast)

designed to work on a strawberry vertical farm. The arms are fixed on the side of the robot and in between the arms, there is an RGBD Zed mini Camera which is used in the perception tasks related to picking and operating the strawberries. The beast robot is controlled using ROS middleware.

7.1.1 Implementation Details

Our perception model is fully developed on Yarp. To operate on different robotic platforms, it is required to implement modules that are capable of acquiring data from the robot and controlling it. For the pepper robot, Four models were developed as the following:

- pepperStartup: a general module that switches on the robot and defines the operating functions.
- pepperImageGrapper: a module that streams images over the yarp network from the 2D camera. It take the camera index as a parameter as well as the framerate.

- pepperAudioGrapper: a module that streams the audio from the desired channels. The module takes the buffer size and the sampling rate as a parameter.

- pepperHeadController: this module controls the hear through Yarp simple commands.

These developed modules are bridging between Yarp and Naoqi SDK. It allows Pepper to leverage in all the functionalities developed originally for iCub using these simple modules.

On the other hand, the beast originally didn't have microphones and the only camera placed on the side for picking has a small field of vision of the surroundings as it is placed only for picking. Therefore we had to add in the hardware of the beast to implement our audio-visual perception model. We added another RGBD Zed mini camera and two identical microphones on the Beast as shown in 7.2. The microphones were fixed facing the front of the robot and the added zed mini camera was placed in between the microphones. Similarly to Pepper two modules for the beast have been implemented as the following.

- beastImageGrapper: a module that streams images over the Yarp network from the Zed mini camera. It streams both the right channel and the left channel.

- beastController: this module is a bridge module between Yarp and Ros. It takes commands from Yarp modules and streams these commands on the Ros network to be executed by the different ROS modules designed to control the robot Including the wheels and the arms.

For the audio, the microphones were connected to the PC on the beast through a mixer. And the regularly developed modules are capable of capturing the audio as it can be read on the PC as an audio device.

7.1.2 Results and Discussion

After implementing the steamers and controller modules for both of the robots, the architecture runs smoothly on both robots without any issues. However, for Pepper robot, the framerate of the images was low compared to iCub and the Beast. Theoretically, it can run up to 30 fps with a low resolution. So it is a trade-off between the resolution and frame rate. In terms of the processing unit, the architecture was running on the laptop of the Beast while for Pepper, the architecture run on a separate machine that are connected with Pepper through the network.

Applying the model in different robots aimed to take the architecture into practicality. We showed that the Yarp Implementation of the architecture can easily be used on other robotic platforms. The only requirement is developing modules for the sensory and action levels that do direct reading and writing functions with the sensors and actuators respectively. This is the first step towards implementing the architecture in real-world scenarios. As real-world applications require a high level of robustness and generalization.

Further steps on this part can be done in two directions. The first direction is to apply this architecture in an end-to-end use case scenario (The following part of the chapter). And the second direction is creating a stand-alone SDK that can be deployed in different hardware operating systems with a simple installation process. This will allow other developmental contributions by different research groups worldwide.

7.2 Use Case: Robot Coordination in a Strawberry Farm

Robots are advancing in agricultural applications. One of the applications is in strawberry vertical farms. Monitoring the state of the berries often takes a big effort. Additionally, field actions with the plants such as dealing with diseases or picking the ready berries is a task that takes a big effort and resources if it not well targeted. One of the possible solutions is to monitor the berries with a flying drone and execute actions on the ground using the Beast. Following this, we saw potential in applying the audio-visual perception model to tackle this application. It is also important to mention that this idea of the application was inspired by the communication of the bees.

In this part of the chapter, we try to answer the second part of the last research question which is to see how the perception model can be deployed in a real-world application. More specifically in a coordination task for the beast with the drone. The task is for the Beast to locate the drone using the audio and visual modalities in the 3D space and then navigate in the farm to reach the position where the drone is flying. This task requires high-level action planning from the beast and perception top-down control to coordinate with the flying drone. The drone might be outside the field of vision which requires the Beast to rotate trying to localize it audio-visually. Additionally, after localizing the drone, the robot needs to navigate the farm to reach the location of the drone.This part of the task requires the robot to coordinate the action and perception cycle to reach the right location actively. Once the robot reaches the desired location, the robot will execute function-related actions such as picking the strawberry (The functional-related actions are outside the scope of my Ph.D.).

We use our architecture which is based on multiple sensories and integrating attention, perception, memory, and cross-modal interaction components to have a complete cycle of perception and coordinated action in a real-world environment. We added new components in the developed perception architecture to have high-level control and edited a few modules to fit the task. In the following section, we will explain more about the motivation of the application and then will explain the implemented architecture. Further, we did an experiment in the lab and a test run on the strawberry farm. More details about the experiment and results will be carried out in the following parts.

7.2.1 Background

Bees Communication Bees are highly social insects. They use various ways of communication between them. These communications are through vision, audio, and chemical signals. They are capable of communicating different kinds of information such as the location of food sources, the status of the hive, the presence of a threat, and even social information to make bonds among each other [92]. They use the auditory channel by producing different auditory signals such as buzzing in low frequency, drumming, clicking, and singing high pitch sounds [139, 113, 90, 92]. Inspired by bees' communication, the idea of a drone communicating information through audio-visual channels to a mobile robot emerged. The drone can fly on the farm and collect information about the state of the farm and the plants and then communicate the required action towards the farm and the plants to the mobile robot using motion gestures. The motion gestures has also an auditory pattern that is possibly detectable by the mobile robot.

From Intelligent Gathering to Action Execution The main use of the drone is to gather intelligence and communicate information to the Beast Robot to act on the ground. The importance of gathering information through a flying drone is to increase the inspection's speed. The drone can fly and move in the farm faster that the Beast. The drone can be programmed to use vision in detecting diseases, check the plant state, and get the location of the rotten berries that are ready to be picked. According to the gathered information, the drone communicates the information to the beast to perform the required action. This communication can be conveyed through flying gestures or a sequence of them. Such as swing, and back flip. The action that the Beast does can be planting a pest or picking the strawberries or spaying a pesticide. This is the main concept of the robot-coordinated task. In this use case, we only consider a simple form of this interaction which is just communicating the location of the drone with a back flip.

7.2.2 Implemented Architecture

The architecture used in this task is part of the whole proposed proactive perception architecture that was proposed in chapter 5 with some added components. It uses the full visual pathway, the localization pathway of the audio, the short-term memory, the multisensory manager, and the action block. The added are High-Level Planning/Task Coordination, Navigation Action Modules, and Navigation Related Long-Term Memory. The high-level planning/Task coordination block was added to manage the task action lifecycle as well as a top-down control of the perception modalities.

As the task of the robot did not need all the functionalities of the architecture, some parts of the architecture were deactivated. These parts are basically the auditory what pathway, the long-term associated experience, and the long-term prior inference in the multisensory manager.

Figure 7.3 shows the cognitive architecture. The deactivated components are shown in light gray color. This version has the high-level planning and task coordination block added, in comparison to the previous version. Also, action modules and a long-term memory component related to navigation. The following part of the chapter will present the functionalities of each block and the added components.

Audio Attention The sound of the drone is usually a low-frequency sound. It depends on the brand of the drone. We tuned the audio attention model to focus only on the low frequencies (<300 Hz) as we found that this is the upper band of the drone which was used in this work. In this version of the architecture, we implemented an adaptive power trigger instead of the trigger that was based on an absolute threshold value that was proposed in Part II of the thesis (specifically chapter 4). The adaptive power threshold is designed to detect big changes in the power level of the environment. It computes the average, maximum, and standard deviation of the power in a defined time window and then defines the threshold based on these values. The implemented adaptive power trigger module takes the triggering mode as a parameter. The triggering mode defines the mechanism of choosing the threshold which can be a fixed added value tp on the average, or to the maximum. It can also be a relative added value (based on the standard deviation) to the average or the maximum.

Audio Localization The audio localization (Where) component is identical to what we used in chapter 5. However, it was tuned to fit and adopt the application (localizing the drone). Tuning the localization system is defining the urgency parameter of the decision-making in the audio manager module. This parameter controls the trade-off between the speed of the

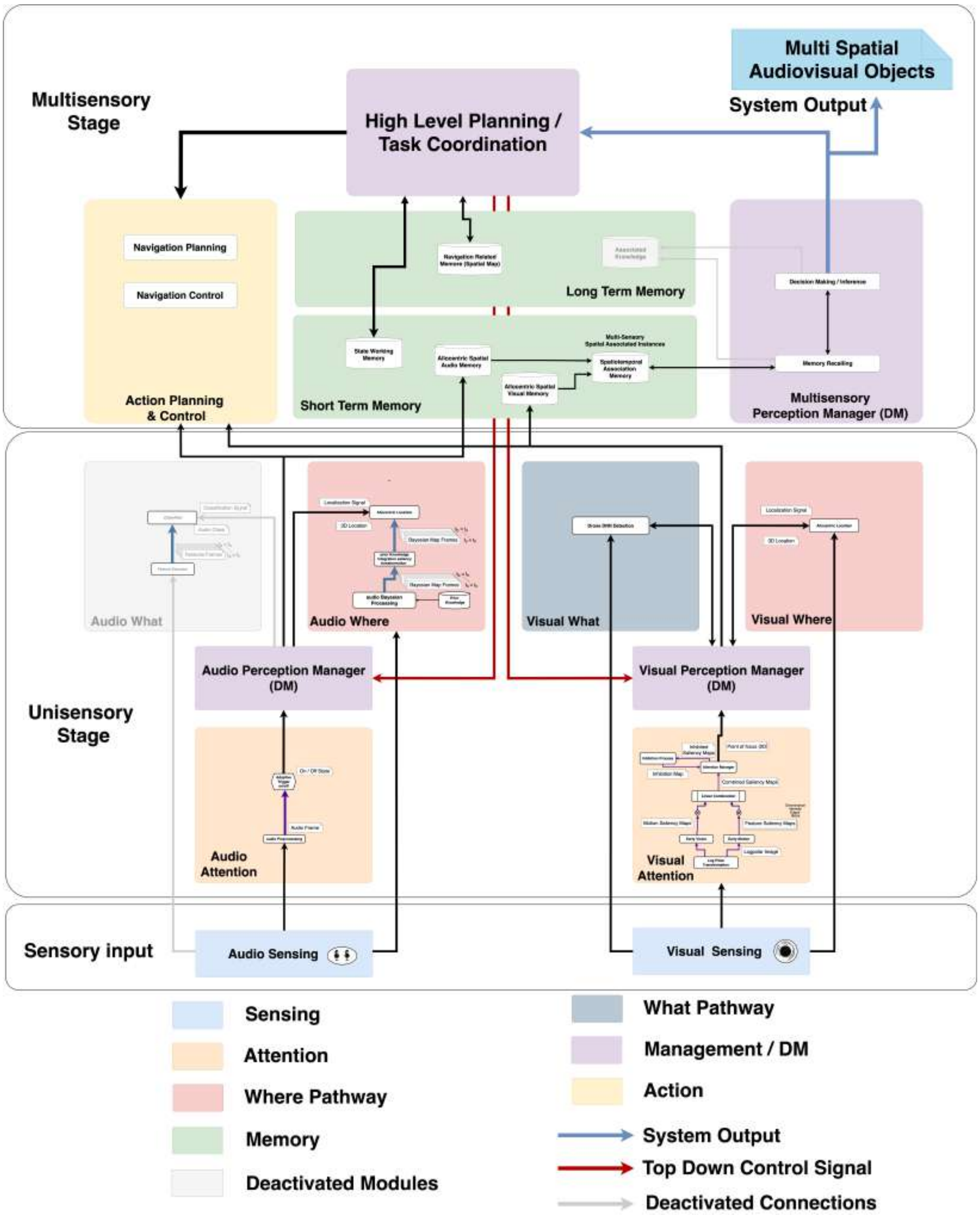


Figure 7.3 The Cognitive Architecture for Action Coordination and Audio-Visual Perception. The components in gray (Audio What, and the associated knowledge in the long-term memory) are deactivated in this version of the architecture. The added components (in comparison to the previous version proposed in chapter 5) are the high-level planning/task coordination block and the navigation-related memory

localization and the accuracy. In this application, the focus was more on accuracy than speed considering the real-world noisy environment.

Visual Attention As the drone is a moving agent, the motion feature is relevant. Therefore, the attention model is tuned to use the motion feature as the most relevant feature. The tuning is top-down control of the model weights. The motion weight was adjusted to one while the rest of the weights were zero. This makes the attention sensitive to the moving elements in the scene.

Visual What For what pathway, we used a pre-trained deep learning model for drone detection [103] and use it in the what module. It perceives a point in the 2D image and then replies with the confidence rate of the existence of the drone in this attended point.

Visual Where The added camera to the robot is Zed-mini Camera. Which provides a real-time point cloud. So the localization algorithm is taking the 2D point as an input and replies with a 3D coordinate. The 3D coordinate is then transferred to the global coordinate system of the robot knowing the exact location of the camera on the robot.

Short-term memory The short-term memory components did not change from the main architecture, except the working memory. The working memory is holding the processing state and the environment states. The processing state is the current state of the processing logic of the task (the logic will be explained in the following part of the chapter). This is an added new state to the original version of the architecture. The environmental states are the states of the visual and audio stimuli which are the main original part of the memory that did not change.

7.2.3 Added Components

To achieve the requirements of the task, a few modules had to be added to our architecture to close the perception-action lifecycle. Basically, the added components are top-level managers and a navigation system.

High-level planning/ Task Coordination Manager This module is processing a high-level control. It does multiple functions. Generally, it is responsible for coordinating the perception systems, the memory, and the action based on the current state of the environment and the robot. It interacts with both visual and auditory perception manages and operates a top-down control for their parameters.

Navigation System The Navigation system consists of two modules and a long-term memory component. The first module is a navigation planner, the second module is a path executor, and the memory component is where the map of the environment is recorded. The navigation planner is responsible for creating a planned path knowing the current position of the robot and the goal in a previously given map (which is recorded in the long-term memory). The given map of the environment is an internal representation of the spatial world using the idea of growing neural gas. The goal-directed navigation is based on the reward fields idea. The origin of this work was proposed by [156] and developed later in [18]. The navigation modules are added as actions in the architecture while the memory component is located in the long-term memory block of the architecture.

7.2.4 Processing Logic

The processing logic is the responsibility of the high-level task-planning component of the architecture. It coordinates between the components of the architecture and controls them as a top-down modulation. Based on the defined task, the logic is defined. In our case, the task is localizing the flying drone using both auditory and visual modalities and navigating toward the localized 3D position.

The system process as shown in figure 7.4 starts with the trigger, which can be happened by a top-down intentional control system or by a bottom-up signal from the environment. The top-down signal is initiated through the high-level control mechanism of the robot which is executed by the operator of the robot. On the other hand, the bottom-up signal is initiated by an auditory event. In our implementation, the bottom-up trigger is initiated from the auditory attention component which is based on the developed adaptive power trigger. The trigger is initiated based on the changes in the power spectrum of the audio. The trigger is on if a high step in power was detected. The step in the power in the auditory environment happens due to the back flip action of the drone.

Once the process is triggered the robot will localize the other agent (the drone) using the auditory where component. The audio localization system is giving an azimuth angle. It is the estimated azimuth angle between the drone and the robot axis. Once the robot has an estimate of the angle, this angle is set to be an action goal to the robot (rotation action). The robot will then execute the action by rotating toward the drone and trying to localize it using both auditory and visual modalities. The localisation and action process is dynamically done in a closed loop based on the error in the estimated position of the robot based on both modalities. The multimodal localization process is processed using the short-term memory and the bayesian integration strategy that was previously explained in chapter 5 which is one of the core components of our architecture. Additionally, the localization processes between the modalities are repeated in a loop until the drone is perceived and localized with enough level of confidence.

Once the drone is localized with enough precision, the 3D location is then set as an action goal. The robot then projects the drone's position to the spatial map to get a target point in the ground. The spatial map is already given to the system and stored in long-term memory. Additionally, the navigation system is tracking the position of the robot on the map. We have to note that the navigation system is only considered an action component in our architecture and it is not the focus of this research. Further, the navigation system plans the path to the goal, and then executes this path, and moves the robot toward the drone. As mentioned in the previous section, the state of the process is saved in the state working memory and managed (updated) by the top-level manager.

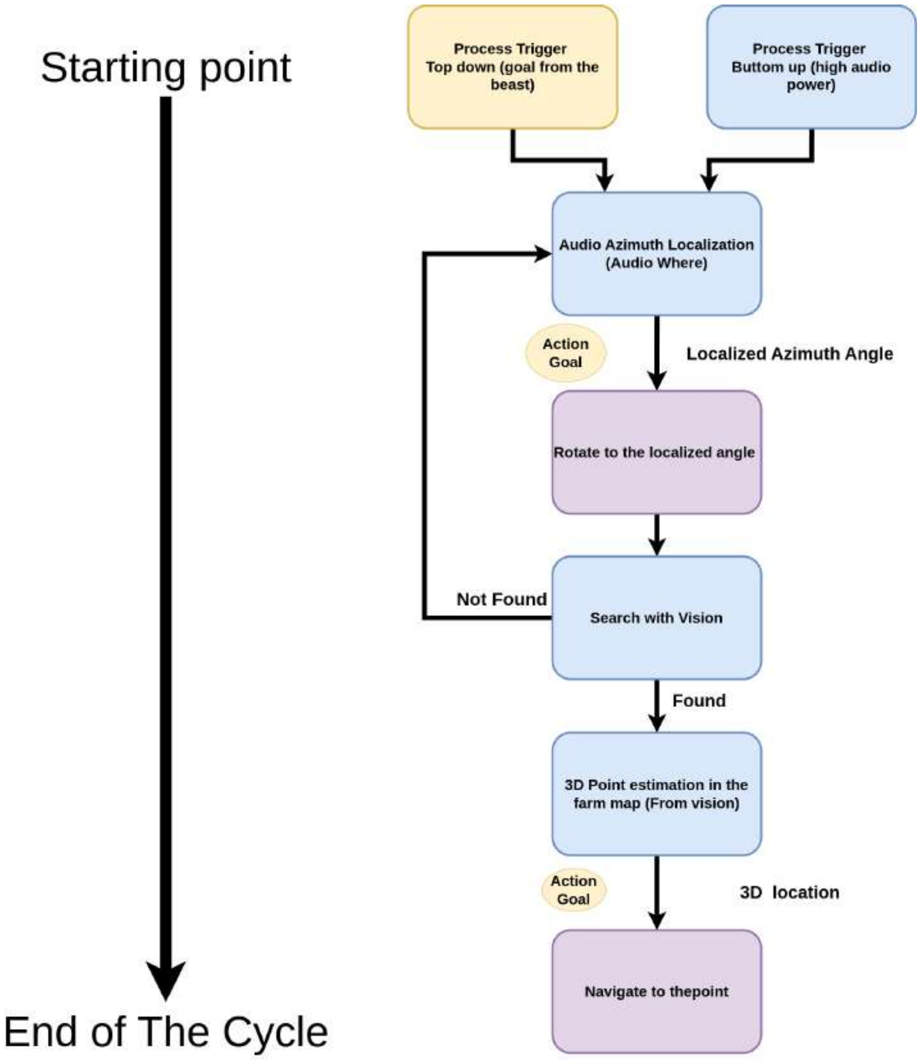

Starting point
Process Trigger
Top down (goal from the beast)
Process Trigger
Buttom up (high audio power)
Audio Azimuth Localization (Audio Where)
Action Goal
Localized Azimuth Angle
Rotate to the localized angle
Not Found
Search with Vision
Found
3D Point estimation in the farm map (From vision)
Action Goal
3D location
Navigate to thepoint
End of The Cycle

Figure 7.4 The Higher-Level Control Process for The Robot Coordination Architecture.

7.2.5 Experiment

To test the functionalities of the system, a lab experiment was made and test runs on the real farm were done. Both the lab experiment and the farm test run were done using the Beast robot. The main aim of the experiment is to validate the overall processes of the architecture and show how can our proposed architecture be used in a real-world application. With this experiment, we would like to move from the abstract experiment in the lab using stimulation boxes that were used in the experiments of the previous chapters toward reacting to real-world stimulations. Applying our proposed architecture to an applied scenario will demonstrate how the architecture can generalize to different tasks.

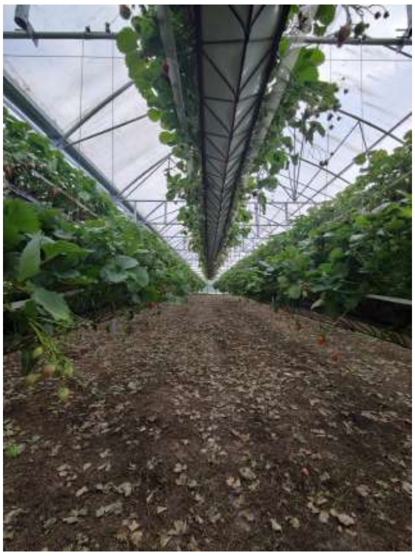


Figure 7.5 Three Vertical Rows in Tiptree Strawberry Farm

Lab Experiment The lab experiment was made to measure the performance of the perception system. The main focus of the experiment is to measure the accuracy and time of the auditory processing, visual processing, and the whole system in general. The lab is equipped with a VICON tracking system. So, we used it to record the ground truth of the drone location and robot location in the 3D space of the lab.

The experiment was flying the drone and moving it randomly in the room in different directions and orientations while recording the response and internal states of the robot. The internal states are decided the location of the drone from the visual modality, and the direction of the drone from the audio modality.

Field Visit The experiment was made on Tiptree farm [1], Essex, UK. The farm is designed as a vertical growing farm. It consists of identical blocks. Each block has east rows and west rows and in the passage in between. Each side of the block has 9 rows. The level of the rows is controlled and adjusted to a specific height. The odd rows are moved together against the even rows (if the odd rows are up, the even are down, and vice versa). Figure 7.5 shows three rows on one side of the Farm. the figure also shows the difference in the height between the middle row and the other two rows as the height was adjusted to raise slightly half of the rows.

[1] https://www.tiptree.com/

7.2.6 Results and Discussion

In this section, the results of both the lab experiment and the test run on the farm will be reported and discussed.

Auditory Localization (lab Experiment) To analyze the robot's performance of the direction estimation, the accepted range of error was defined to be 5 degrees as it is a close value to human capabilities. If the error of the estimated azimuth angle was more than 5 degrees, the estimation is considered wrong. Following this rule, the applied architecture on the Beast robot achieved 71% accuracy in the lab experiment.

Visual Perception (lab Experiment) For visual perception, It is divided into three processing, the first one is attention followed by both 3D localization (Visual Where) and drone recognition (Visual What) blocks. After these blocks, the visual perception manager block decides the final estimate of the localized drone.

To evaluate the visual perception part of the architecture, we compare three different strategies for the visual perception task (Localizing the one). The first strategy is basing the 3D localization based on the attention model only. The second strategy is using the deep neural network model of drone detection[103]. And finally, the third strategy is using the proposed model which is based on attention, and the dual visual pathways to decide the right location of the drone in space.

Figure 7.6 showed the accuracy of these strategies. It shows that using the proposed model for perception led to more accurate decisions (around 75%). This accuracy is higher than both the attention-only strategy (43%) for attention, and (60%) for the DNN algorithm.

Field Visit Observations In the Field, there were some observations as the examination was qualitatively made.

The audio environment of the farm is very noisy, because of, the workers, fans, cars, and tractors working close to the field. However, the environment has fewer reverberations in comparison with the lab as the farm can be considered an open space. The auditory attention system has to be tuned to orient the focus toward the frequencies that the drone produces. After the tuning process, the robot was having a fairly good response to the sound of the drone.

For the visual part, the drone detection deep learning algorithm was detecting some leaves as a drone, however, the attention system helped to regulate this issue by only focusing on the moving areas of the environment. As the motion feature had the highest weight in the

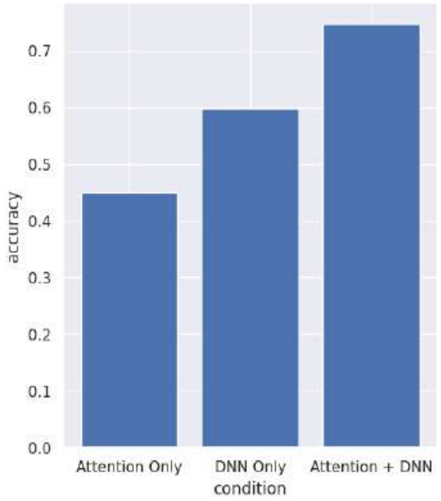


Figure 7.6 Visual localization and detection accuracy using different methods on the Beast robot in the lab experiment

linear combination system by design. Another observed point was the latency in the visual processing affected the 3D localization.

Generally, the whole system was functioning and the robot did a complete cycle for the process properly. The robot was capable to localise the drone actively using both audio and visual channels. Further, the 3D point of the localized drone is projected on the map of the environment (the farm). This projected point is then used to execute the navigation action toward this location.

Discussion The used drone in this experiment (Dji Tello mini) is relatively a small drone. localizing the drone is actually a challenging task using visual modality, audio modality, or even both of them. On the farm, the distance of the drone from the beast was within 10m distance while in the lab was within 5m considering the available space of the lab. At the beginning of the testing trials, the drone was flown outside the Beast's field of vision. The robot has to decide the azimuth angle of the location of the drone and the direction of the rotation to ensure that the drone is in the field of vision. Further, decide the location of the drone audio-visually. Then plan the navigation path and execute it.

Although the auditory environment on the farm was very noisy, it had the advantage of non-reverberation (open fields). This advantage allows the robot to localize the sound faster than in the lab experiment. Surprisingly, we observed that the auditory attention after tuning worked pretty well to filter the other noise (such as speech). This is mainly due to the low frequency of the sound generated by the drone in comparison to the speech and other sounds in the environment. This demonstrates that sound filtering might be one of the solutions to the noisy environment. Indeed, we used a simple filtering technique, but more advanced methodologies for selection and filtering are key skills in auditory processing in robotics. In fact, dynamic methodologies will be the gate for solving complex auditory scenes such as the cocktail party problem [81].

The 3D localization in real time for a moving object (the drone) was having a bit of latency. This is mainly due to the limitations of the processing power and the hardware technologies. This latency can be solved by running the drone localization model on a pc with a GPU, or using lightweight models for drone localization that has a lower processing time. This is actually one of the robotics challenges while using deep architectures that require a big amount of memory to save the parameters of the model, and the processing power to reach high frame rates. In chapter 9, we will explain more about this challenge and propose some alternative methods that might suit robotics applications.

Overall, the architecture efficiently was capable of performing the task. The robot was capable of initially localizing the drone through the audio only, orienting itself towards the localized angle and simultaneously localizing the drone through both audio and visual channels for more accurate positioning, and finally navigating towards the localized 3D position in the farm. The results showed that using the architectures method in the visual perception task (combining attention with the dual pathways) improved the accuracy of localizing the drone in comparison with relying only on using the neural network model.

The architecture was used to coordinate the robot's actions with the behavior of another robotic agent through audio-visual perception. This doesn't imply human-robot interaction. Although, the experiment is to show the capability of achieving a coordinated task with another agent in general. Using an artificial agent while testing allows more flexibility and eases the developmental and validation process.

7.3 Conclusion

Building a generalized architecture for robotic perception is one of the current challenges. In the previous chapters, an audio-visual perception architecture was proposed. This chapter demonstrates how the proposed architecture was applied on two different robotic platforms other than iCub which are The Essex Agricultural robot (The Beast) and Pepper. For Pepper, we only adopted the software implementation to show that our software is not specific for iCub (the original robotic platform that this architecture was developed on). This will help the research community to further develop and use our software (We will make it available for the research community). For the Beast we take it one step forward after adopting the software, we applied the architecture in a real-world scenario using a sub-part of the architecture with some added components to do high-level planning and coordination. We used the architecture in an agricultural application. The application is coordinating between the Beast and a flying drone. The drone is responsible for gathering information about the farm and trees and communicating this information audio-visually with the Beast (By doing a backflip). The architecture was applied on the Beast to localize and detect the drone and move towards the projection of the location where the drone was detected. The added components to the architecture were the connection between perception, high-level arrangement, and action. The architecture was tested in a lab environment and in a real Farm (Tiptree, Essex, UK). The results showed that our proposed architecture can be technically valid to be applied in a real-world application.

Additionally, this development of the application shows the modularity and expandability of the architecture because of the following:

- Using a sub-part of the architecture
- Adding action and task-specific processing
- Tuning the modules of the robot to fit the different hardware (robots) and the application

With this research work, we tackled the last research question *RQ3* which is about generalization and coordination in a real-world scenario. We applied the model in different robotic platforms by adopting the inputs and outputs of the architecture to fit the robot. High-level coordination was enabled in the architecture by adding a high-level manager that controls the perception and action flow. Additionally, it controls the perception modalities as a top-down behavior. The experiments done on the architecture showed the main functionalities of the architecture. However, the architecture is very adaptive and modular which makes it an effective tool to be used in robotics in many activities including human-robot interaction activities.

In this part of the thesis (Chapter 5, 6, and 7), we proposed an audio-visual cognitive architecture for robotic perception. Our contributions to the community with this architecture are as the following:

- The architecture is addressing **cross-modal interaction** in which the previously perceived associated features from different modalities are influencing the localization and identification process. This is the first perception cognitive architecture that so far addresses this gap, to the best of our knowledge, and as explicitly mentioned in the latest review of cognitive architecture [120].

- The architecture is biologically inspired and implements the perceptual inference with the consideration of the prior knowledge in different perceptual levels. Alongside perceptual inference, it adopts the organizational structure of the perceptual information in the cognitive components such as dual pathways and spatial working memory. It also considers high-level planning and task coordination which allow the robot to actively interact in the environment with other agents and coordinate their work together to achieve the goals of the required task

- The contribution wasn't only with the theoretical idea, but we also implemented this architecture in multiple robots and demonstrated that the architecture can be used in real-world applications. We applied this architecture on three different robotic platforms which are iCub in the main experiment in Chapter 5, and Pepper and the Beast in the current chapter. We showed that the architecture can be applied to different robotic platforms and applications by only tuning the parameters and using task-specific components such as the drone localizer that replaces what component for the vision as well as the audio. which makes the proposed architecture a general-purpose perception-action architecture for different robots unlike the existing architectures that are specific to one task (e.g. speaker localization)

- Across the development phases, we show the modularity and expandability of the software implementation. In particular, in chapter 6 we used the auditory perception component with the action component to validate the active behavior of the architecture and propose a methodology to examine different active motor behaviors to improve the auditory perception. While in this chapter (chapter 7 we disabled the auditory what pathway and associated knowledge in the long-term memory while keeping the rest of the architecture with added high-level planning component.

Considering this point, we present this architecture as a base implementation towards audio-visual shared perception. The evaluation of this architecture mainly examined the capability of the architecture of solving complex scenes. The next evaluation step is to compare the behavior of the architecture with the human perceptual behavior. Aiming to examine if the architecture archives the same perceptual characteristics including common representation and spatiotemporal coordination which are two of the five shared perception skill categories that were framed in the first two chapters of the thesis (chapter 1 and 2).

With the proposed architecture in this part, we tried to address an architecture that shares a common environmental representation and proactive behavior (common representation skill category) and high-level coordination and planning (spatiotemporal coordination skill category).

In the following part of the thesis, I will present the work done to improve the auditory pathways. This is mainly because of the observed limitations in the auditory modality.

Part IV

Improving Audio Pathways: Alternative Pipelines For Developing Robot’s Audition

Chapter 8

Audio HRTF: An Alternative Pipeline for Sound Source Localization in Robotics

Mr Watson come here i want to see you

Alexander Graham Bell in first words ever spoken on the telephone

In the previous chapters, I demonstrated the developmental process of my perception architecture. The bottleneck of the process that limits the speed/accuracy of the robot was mainly the auditory localization model. This is generally due to the ego noise and electronic noise of the robotic environment. Furthermore, the challenges even increase in more complex scenarios in human-robot interaction such as cocktail-party, multi-speakers, and multi-agent interactions. Trying to fulfill this gap, this work aims to develop better auditory building blocks for robots. More specifically, is to improve the speed and accuracy, address multiple sound sources/speakers existing in the environment at the same time, and overcome noise and reverberation in the environment. In this chapter, we propose a developmental pipeline for robots using Head-related transfer functions (HRTFs). The pipeline starts with the measurement of the HRTFs for the robot, following the creation of a specialized dataset, and finally training a model using this dataset. We applied this pipeline to develop a single-speaker localization in the azimuth front field. The performance of the model showed promising results that exceeded the current models for robots. Further, the aims are to increase the dimensionality (including the distance and elevation), render more datasets that simulate the ego noise and multiple speakers, train more models using these datasets, and finally test these models in the real robot. Additionally, this pipeline can be a tool to analyze and understand different robotic auditory contexts that will hopefully help in developing better model-based architectures and systems for robots.

8.1 Background

8.1.1 Human Sound Source localization

Sound source localization is the ability to identify the direction where the sound is coming from. The study of sound localization is focusing on both absolute and relative localization. Absolute localization is judging the absolute position of the sound source in the 3D spatial world. On the other hand, relative localization is identified by the minimum audible angle (MAA) which is the minimum detectible angular shift between two sources or locations in space.

Humans are capable of localizing the sound source using multiple monaural and binaural cues. The monaural cues are the cues that are due to the interaction between the sound source and the environment through which it propagates, notably including the physical anatomy of the head and the ears. This interaction is the effect of the anatomy of the head and pinnae on the sound signal before the signal enters the canal of each ear, and has the effect of spectrally filtering the sound signal. By contrast, binaural cues arise due to the differences in the sound between the ears. The most common binaural cues are interaural time differences (ITD) and interaural level differences (ILD). As sound waves travel in space at the same speed and the ears are in different locations, the arrival time of the sound for each ear is different. This difference is what ITD refers to. It can be represented as a delay between the ears which depends mainly on the frequency of the sound. Similarly, the ILD refers to the difference in the intensity level of the sound from one ear to another. This happens because of the propagation of sound in space. When the sound travel in space, the energy of the wave decrease due to the increased area where the sound is spread. Therefore, the intensity level of the sound is higher for the closer ear from the sound source. And this is what the brain is resolving to measure the sound direction in the horizontal space.

Although the binaural cues are significantly important in sound localization, they can only solve azimuth sound problems and they face the problem of the cone of confusion [22]. This is the cone-shaped set of points in the space around the listener where the sound source might be located. As all of the sound sources located in these points produce the same binaural effect, there is ambiguity inherent in both ITD and ILD localization. This adds a challenge to the binaural sound localization models. One possible solution to the ambiguity in the egocentric and allocentric spaces is including the head movements in the computation process [79].

In addition to the azimuthal cues provided by ITD and ILD, the asymmetrical shape of the pinna acts as a filter that deforms the sound signal and imparts a spectral filtering characteristic that is dependent on the vertical position of the sound. This further helps to distinguish the right source of sound [15].

One common approach that computationally describes the characteristics of the auditory signal modulation due to the physical shape of the head/ears is the head-related transfer functions (HRTFs). The HRTF characterizes how a sound that comes from a point in space will be modulated before arriving at the inner ear. The HRTFs are holding both the binaural and monaural cues in the characteristics of the functions. Therefore they are very specified and sensitive to changes in the shape of the head/pinna. By emulating the transfer function of free-field sound arriving at the inner ear, the HRTFs are widely used to create (or render) artificial auditory spatial experiences through headphones. This is often exploited in presentation media such as in augmented reality. Although they allow the creation of directional sound experiences that are very close to real-world experiences, a non-trivial challenge to obtaining optimal sound rendering is the complexity of the measurement process itself. Ideally, a custom listener-specific HRTF is used to render virtual auditory space on headphones.

In addition to virtual auditory displays, HRTFs have an important usefulness in spatial audio research. In many situations, it is desirable to be able to programmatically and reproducibly render spatial sound cues, for example when training an artificial neural network to extract spatial features. Displaying an auditory scene in a free field from actual loudspeakers would require a very long time and require dedicated hardware and an acoustically controlled space. Instead, large training sets can be rendered *in silico* using HRTFs and the resulting left and right channels (ears) can be used instead of capturing true free-field audio.

8.1.2 Sound Source localization in Artificial systems

In artificial systems and robotics, much attention has been placed on machine vision systems, however, the artificial and robotic audition is relatively less explored. The research in sound source localization for robotics is often found in signal processing and robotics communities. In the signal processing and robotics communities, researchres often build different models that localise the sound source in the azimuth direction and distance. The azimuth angle has taken most of the focus compared to the distance and elevation because it is the most informative dimension in human-robot interaction. In recent years, many models have been developed using an array of microphones. Fewer models used humanoid binaural audio (only 2 microphones) in sound source localization [136]. [8, 197] are two recent reviews for sound source localization in robotics. There is interest in using binaural audio in the robotics community because of the efforts to minimize the number of sensors that are directly affecting the energy efficiency of the system, as well as representing a human-like audition. The binaural audio models are generally divided into two different categories, the first category is using deep-learning models whether supervised or self-supervised, while the second category is implementing computational models for the biological audition systems. The advantage of biologically inspired modeling is that it doesn't need data to train like deep learning models. So it is easier to implement in different platforms and different environments. However, these models generally struggle to deal with noise, and it is difficult to achieve high performance using these models. On the other hand, deep learning models are achieving better performance and are widely used in the signal processing community but acquire a huge amount of labeled data for robust performance.

Going further details on recent research in robotics, [70] worked on developing a self-supervised deep learning model to acquire a dataset and train a deep neural network. The network was trained to localize a single speaker. The dataset is acquired through an interaction with the robot in which the robot localises the speaker visually and uses this localization to annotate the auditory data. His model achieved good accuracy but for a large spatial resolution which is 3 or 5 bins for the front field. Similarly, [83] developed a deep learning model with manual annotation for the data and used the four microphones that the Pepper robot has.

8.1.3 Challenges in Robotic Sound Source Localization

In a recent review of sound localization methods in robotics [197], they highlighted some challenges in this domain. Dealing with the dynamic acoustic and complex environment, and having spatial sound datasets are two of these challenges.

In many human-robot interaction scenarios, the robot is interacting with multiple agents at the same time. Additionally, the environment has different acoustic characteristics and might also have auditory objects such as speakers and buzzers. In general, robots work in dynamic acoustic and complex environments. Therefore, multi-agent audio localization models are a requirement for real interactive scenarios whether this agent is a human (speech) or something from the environment (environmental sound). More specifically, it is important for the robot to have the skill to localize different kinds of audio signals while other audio signals do exist in the scene or some even more complex situations such as the cocktail party scenario.

The second challenge is the availability of labeled spatial data that can be used to train different models to achieve different robot audition tasks such as localization, multi-speaker localization, and noise cancellation. In particular, the available datasets do not replicate the robotic environmental conditions including the ego-noise, reverberation of the sound in the room, and generally the interaction dynamic conditions. The other use of spatial data is the evaluation of the model-based approaches. This is due to the difficulty of getting accurate ground truth (in degrees) considering the movements of the robot and the humans in the scenario.

8.1.4 Motivation

We mentioned that in our architecture, the bayesian localization model which is based on the interaural time difference was generally the bottleneck of the process. It was relatively slow when trying to achieve an acceptable confidence rate (accuracy). This was mainly because of the effect of the ego-noise which is due to the fan and the electrical noise in the head of iCub. In fact, the model wasn't designed to be fast but was designed to integrate spatial evidence over long periods of time to build a probabilistic map for the scene.

On the other hand, the previous section of this chapter showed some of the challenges in SSL in robotics.

In the literature, there is not much work yet done in the direction of using the HRTFs of robots to develop spatial hearing. In [88], they overcome the computation complexity of the HRTFs by creating a close estimate of the physical shape of the head. However, this method can't be generalized to be used in different robots considering the different head shapes of the robotic platforms. [109] showed the potentiality of using the HRTFs and implemented a localization system based on the HRTFs on the KEMAR manikin with modified human-shaped ears. However, the system is very specific to the designed environment in which the HRTFs were measured. If there was any change in the environment or the head has slightly moved, the system performance will decay.

To fulfill this gap and overcome the challenges, we aimed for two goals. The first one was to improve the localization performance using the HRTFs of the robot as an alternative to the bayesian localization model, and the second goal was to explore the possibilities to use the

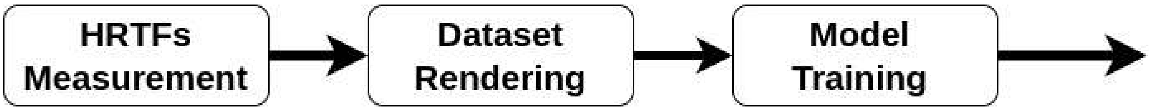


Figure 8.1 iCub HRTFs Audio Pipeline

HRTFs to create models that solve complex auditory scenarios in human-robot interaction such as multi-speaker localization.

8.2 Methodologies and Experiment

I propose here a developmental pipeline for sound source localization in robotics. The pipeline is using the HRTFs to render specialized datasets, and then trains different localization models using those HRTFs and adapted to specific required tasks. This is mainly solving one of the mentioned challenges for SSL in robotics which is the availability of spatial audio datasets. Using the HRTFs, it is straightforward to render a spatial dataset personalized to the robot. Additionally, this will further help in addressing another challenge which is the acoustic dynamic and complex environment of robots.

The pipeline is divided into three stages, the first stage is to measure the HRTFs for the robot. The next stage involves the rendering of specialized spatial datasets. And finally, training models are created to address the auditory task. Figure 8.1 is showing the pipeline process. In the following part of the chapter, I will explain the process of each stage and also what was done as a demonstration and validation of this process. The demonstration and validation objective was to build a single-speaker localization model.

8.2.1 Stage One: HRTFs Measurements

The first stage in the pipeline is to measure the HRTFs of the robot. There are various methods that are commonly used to measure HRTFs. In [135] the authors reviewed the measurement methods of the HRTFs. The HRTFs are commonly used in virtual and augmented reality to create more personalized experiences, however, it is less common in the robotics field due to the complexity and time of the measurement process. The measurement of the HRTF for any particular robot generally involves recording the unique impulse responses of the head. The recordings are made at various positions (angles) with respect to the sound source and usually in a non-reverberant environment. The impulse responses are recorded using different kinds of excitation signals such as random noise signals, stepped-sine signals, and sweep signals. Further, the impulse responses and the original excitation signal are used to

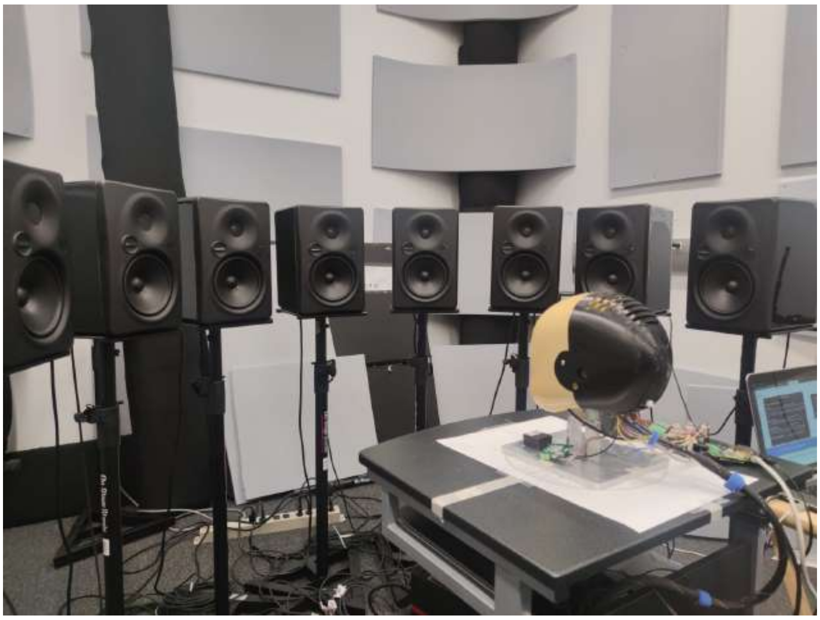


Figure 8.2 Impulse Responses Measurement Setup

compute the HRTFs. The computation of the HRTFs from the impulse responses is generally made using several methods in the frequency domain or the time domain or using the least squares fitting method.

The Setup The HRTFs measurement process was made at Tata Lab in the University of Lethbridge, Canada during my second period abroad. The lab is specialized for human auditory perception and robotics research. It has a minimally-reverberant room, free field studio monitor speakers, and an iCub head. The figure, 8.2 shows the setup of the microphone response measurements for the iCub head. The iCub head was placed in front of the middle speaker which was the only one used in the process. To obtain high-quality audio recordings, a Reverb Robotics ReRo board with digital MEMS microphones was fitted to the iCub head. This is a small-sized integrated board specialized for auditory perception in robotics research [1].

Recordings An exponential sine sweep signal was used as an excitation signal using 20Hz as a starting frequency and 24000Hz as the maximum frequency over 20 seconds duration. This signal is particularly useful due to its characteristics as the following:

- It covers a wide range of frequencies, which makes the measurement suitable for different kinds of audio signals.

[1] https://reverbrobotics.ca/

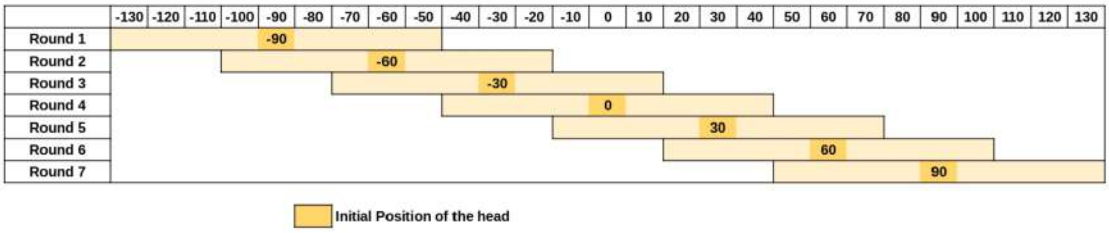


Figure 8.3 Summary of recording angles

- It has constant energy and time which ease the calculation of the HRTFs, and eases the interpretation and analysis of the recorded signal compared to other signals that have variation in energy and time.
- It also has a good frequency resolution, which will reflect of the accuracy of the HRTFs measurement.

As a first step, we only focused on the azimuth angle in the front field with a fixed distance. The iCub head has 40 degrees range of motion in yaw for each side (right, left). I recorded seven rounds. In each round, the robot was placed at an initial angle with respect to the middle speaker. (-90, -60, -30, 0, 30, 60, 90). In each round the robot was moving from the maximum left (-40) to the maximum right (40) with a one-degree step size. For each position of the head (degree), the excitation signal was played and the response of the speakers was recorded. This covered azimuth degrees from -130 to 130 which covers the whole front field and a small part of the backfield with some repetitions of the same azimuth angle. The summary of the recordings angles is shown in Figure 8.3. From the software point of view, a system was developed using Yarp and Rero software [2] to simultaneously record and control the robot's head motions. The system recorded both the microphone signals and the head state from the encoders of the motors in the head.

HRTFs / HRIRs computation from the recordings The HRTFs can be generally obtained by applying linear deconvolution on the recorded signal with the excitation signal in the frequency domain. But the deconvolution process is complex in time. Instead, it is easier to compute its representation in the time domain which is called Head Related Impulse Response (HRIR). The HRIR is computed by applying a convolution on the recorded signal with the inverse filter of the excitation signal. The inverse filter of the exponential sine sweep

[2] https://github.com/reverbrobotics/rero_core_dist

signal is the reversed and normalized version of the same signal [55]. We followed this method and then saved the response in both the time domain (HRIRs) and the frequency domain (HRTFs).

8.2.2 Stage Two: Rendering a speech spatial dataset

Creating a dataset for a specific task is the second stage after the measurement and computation of the HRTFs. Recording such a dataset is generally a time-consuming and difficult task. However, using the HRTFs to render a spatial dataset facilitate creating different kind of binaural datasets for the robot using the existing single-channel available audio datasets. This is done by applying convolution of the signal with the HRIR in the time domain or multiplication of the signal with the HRTF in the frequency domain.

To focus on human-robot interaction scenarios, I created a spatial speech dataset for iCub using the recorded HRIRs and LibriSpeech dataset [181]. LibriSpeech is a speech dataset which is derived from audiobooks. We used the 100 hours "clean" speech version to generate 500 hours of spatial speech. Each audio sample is rendered in 5 different angles assigned randomly between -90 to 90 (the front field). The audio files of Librispeech have variable lengths.

8.2.3 Stage Three: Model Training and Evaluation

The final stage in the pipeline is training a model using the rendered dataset. In recent years, many models have been proposed to solve sound localization, detection, and tracking algorithms using deep learning [77]. Therefore, we used here ResNet50 [82] architecture as a baseline model in deep learning. We trained a ResNet50 network with modified input and output layers to match our data and task. We used randomly selected one-second clips of audio for each training step. The inputs of the network were the magnitude and the phase components of the short-time Fourier transform (STFT) of the audio signal, which is widely used in the literature. We annotated the rendered data set with 5 degrees of spatial resolution in different segments. This results in 36 segments for the rendered dataset. Thus, the output layer of the network is a linear layer with 36 neurons. The segment is determined by applying a softmax function on the output layer. Figure 8.4 visualize in brief the localization model. For the implementation side, rendering the dataset, and training the model was done in python using NumPy, PyTorch, and PyTorch lighting libraries. We split the dataset into training (70%) and testing (30%) splits.

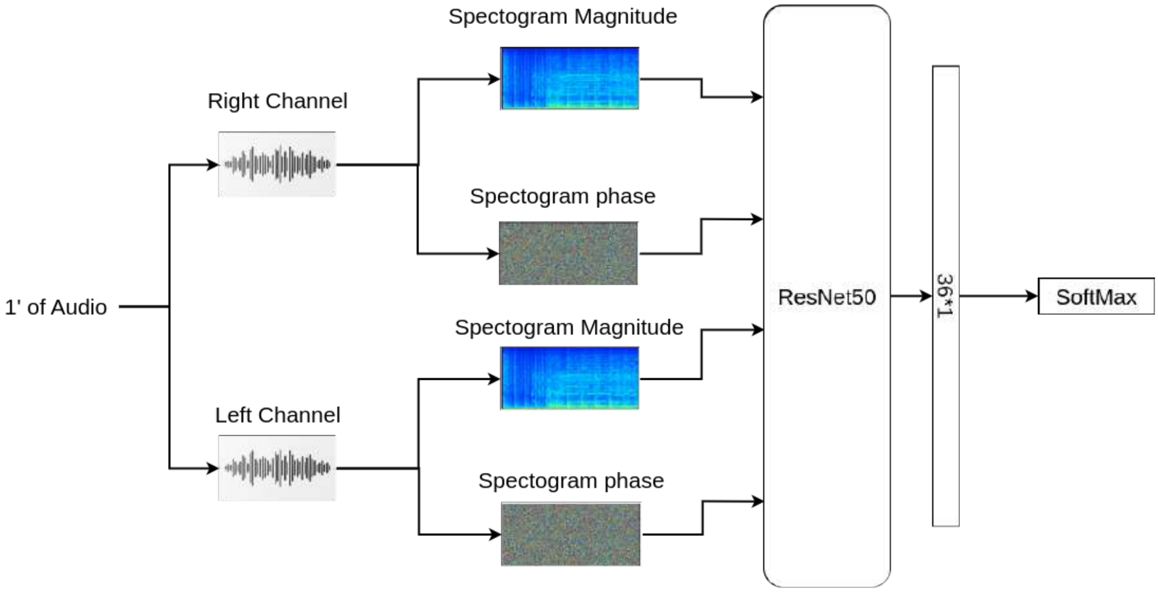


Figure 8.4 The Model for SSL (Single Speaker)

8.3 Results

The training resulted in 94% testing accuracy and 99% training accuracy using one second of audio. Indeed, these results are on the generated dataset. It shows that the trained model can capture the characteristics of the sound signal using a short period (one second), with a 5 degrees resolution. The next steps of the evaluation which we aim to perform in the near future is to evaluate the trained model using the dataset that was recorded using iCub (with the full body) and used to train a deep-learning localization model to identify the source from one of 5 directions (far left, middle left, center, middle right, and far right) [70]. This will give us more insights into the performance of our approach with the comparison of the baseline.

This is the experiment done Following the development pipeline. It showed that using this pipeline is useful in making effective models for robotics. The network was able to capture the characteristics of the HRTFs and learn the first task (single speaker localization). Indeed, the next step is deploying this model in the robot and testing it in different scenarios such as reverberant/non-reverberant environments, or different versions of the iCub (iCub Head, the full iCub) and comparing the performances. For the task of single-speaker localization,

8.4 Discussion and Conclusion

Sound Source Localization (SSL) is an important building block in the robotics domain. Different challenges in this domain were demonstrated such as the dynamic and complex environmental conditions and the availability of specialized data for robotic applications. In this chapter, I proposed a developmental pipeline for the SSL based on the Head Related Transfer Functions (HRTFs) for robots. The pipeline goes through three main stages: The first stage is measuring the HRTFs of the robot, the second step is rendering a specialized dataset, and the third step is building and training a model for the task. As a first step in the SSL development using this pipeline, I measured the HRTFs for the iCub robot in the azimuth front field. Then, I rendered a spatial speech dataset using LibriSpeech which is one of the public datasets for speech. Finally, I used this dataset to train a classifier to localize the speaker. The classifier model was a RESNET50 neural network that takes one second of audio as input. The classifier showed robust and accurate testing results (94 %) with a 5 degrees resolution. This resolution and accuracy are very comparable with human capabilities. The results validate the approach and open up a new line of research using this pipeline. This approach is showing a very strong potential to solve two of the challenges in SSL for robotics. Rendering specialized datasets using the measured HRTFs is a solution for the limited availability of datasets that are specialized for robotic applications and environments. Additionally, this will give the flexibility to simulate different scenarios such as the cocktail party, reverberation, and ego noise. The simulated scenarios can be further analyzed to understand more about the complex acoustic environments in robotics and build suitable systems whether model-based and inspired by biological hearing or data-driven models. The future plan is to make iCub HRTFs and the HRTFs development software package available for the iCub research community in partnership with Reverb Robotics. This will be through the integration of this work into Rero software which is available under a free license. This way will prevent other labs to reproduce the pipeline and can directly proceed in training models for different tasks. Additionally, the datasets will be also published for the whole robotics community. This will allow other robotics communities to use the development software as well as the datasets so that they do not have to spend time on the software code and only focus on running the software on their robots or training their preferred models using the datasets.

Chapter 9

Audio Shallow Models and Continuous Learning: Alternatives for Deep Learning Approaches

Intellectual growth should commence at birth
and cease only at death.

Albert Einstein

In our proposed audio-visual perception model (Part III of the thesis), we trained an auditory classifier which was presenting the "what" pathway for audio. The classifier was trained on four complex tones. This task is relatively easy and did not take time to be trained as we used the same exact signal in our experiment. However, this is not the case in real interaction. For example, in the applied scenario in the strawberry farm, the next step was communicating different kinds of information through different motion behaviors for the drone. Real scenarios often have a more complex scenario than the experiment we have done. Therefore, a more advanced sound classification model is required. In this chapter, we explain our research to explore shallow models and continue learning for audio classification.

9.1 Introduction

In recent years, deep neural network models became one of the most used approaches in intelligent systems [106]. This is due to that it achieved higher performances compared to other traditional approaches in computer vision [19, 99], and language processing [56]. These advancements helped in designing and performing human-robot interaction experiments.

However, the dependency on data, the complexity of both time and space (Training, and Storing the model), and the required computational power of this approach put a challenge in the development of human-robot interactions in real scenarios. The small size of the available datasets for human-robot interaction [24], and the limitation in hardware (use of the CPU only on the robot) are the main reasons.

One of the solutions is transfer learning, which is using a pre-trained model and only train the last layer to fit the required task. Some popular networks are VGG16 [227] or ResNet [82] (vision), respectively, VGGish [85] or YAMNET[1] (audio). Still, the required data to create a generalized model that fit different interactive scenarios is comparably big.

One another approach is developmental learning. It originated from the inspiration of perceptual development in infants. Instead of learning many classes once, the agent learns incrementally by experience and exploration [218, 258]. It is also called continuous learning (CL). In this context, the data is presented as a stream without prior knowledge about the classes. The CL approach is associated with *catastrophic forgetting* [62]. It is the main limitation of CL approaches that has been addressed by proposing different kinds of regularization mechanisms.[114, 207, 245, 208, 183, 243, 213, 184, 256]. Further information about CL can be found in these two recent reviews [132, 47].

One common approach to take the advantage of the convolutional neural networks (CNN) in audio is to transform the raw audio to its spectrograms or Mel spectrum and deal with this as an image. This allows the use of the CNN pre-trained networks like in [85, 78]. However as mentioned before, this was using closed-set annotated data which is not always the available case in HRI scenarios. Therefore, continuous learning and unsupervised learning might suit the robot's scenarios where the robot is capable of acquiring experiences continuously during the interaction and learning using this data. This can be applied for different auditory tasks like audio segmentation [146, 35] speaker identification [5, 97] or acoustic emotion recognition [54, 171, 134]. I focus on the relevant task to our developed model which is sound classification. The unsupervised learning models are usually the focus of the lightweight research line such as in [153, 170, 61]. In there, authors use different approaches such as few-shot learning, zero-shot learning, and open-set classification. In the robot audition community, most of the focus was on localization, separation, and speech recognition [177]. However, very few works have been done in the direction of environmental sound classification and recognition. In [254, 169, 112], the authors proposed a method for noise-robust feature extraction that considers spatial separation using HARK [168] (open-source software for robot audition). In [209], authors propose software for detecting environmental sounds.

[1] https://www.tensorflow.org/hub/tutorials/yamnet

Recently, a convolutional neural network was proposed to perform sound localization and classification on robots [23] but it was done in simulation (not on an actual robot).

The aim of this research part is to explore alternative methodologies for audio classification for robots. I want to close the gap between machine learning and human-robot interaction. I used the ESC10 and ESC50 audio classification benchmarks [187], as well as a modified dataset extracted from the 'dcase19' sound challenge [2] in our study. We conduct experiments using the popular VGGish network as a performance upper bound and contrast the results with the k-nearest neighbor algorithm (knn), which has been also a baseline for the ESC10 and ESC50 dataset thus it serves as a representative unsupervised model. Also, we employ an echo state network (ESN) [96], a recurrent neural network model whose training contracts to a regression task and is, hence, faster than gradient-based training. This qualifies ESNs for robotic applications [86, 80], however, studies on their relevance for audio tasks (specifically within robotic settings) are still sparse [100].

9.2 Datasets, Preprocessing, and Audio Representations

As the main focus of my research is understanding the audio-visual environment, I focused on datasets from real-world environmental sounds such as animals, alarms, ..etc. Four classes have been selected from 'dcase2019' audio benchmark[3], namely 'alarm', 'blender', 'dog bark', and 'dish washer' as the experimental setup used in the research involves four audiovisual objects. A small dataset has been recorded to emulate the audio quality arriving at the iCub microphones. This dataset will be referred to in the thesis as 'dcase_icub'. To understand the scalability of the used models from the number of classes perspective, ESC10, and ESC50 [4] (Environmental Sound Classification) dataset[5] have been used. ESC dataset has a wide range of classes for nature sounds (e.g. ran, baby cry, telephone, etc.). Table 9.1. shows the summary of the datasets.

For the preprocessing part, the normalized Mel spectrum and the corresponding Mel cepstrum coefficients have been computed from one channel of the audio signals (for simplicity). The coefficients are a compact representation of the Mel spectrum and are obtained via a discrete cosine transformation from the Mel log scale. After preprocessing, high inter-class variability was found as shown in figures 9.1-9.4 for the dcase_icub dataset. This is apart from the expected variability between different classes.

[2] https://dcase.community/challenge2019/index

[3] http://dcase.community/challenge2017/task-rare-sound-event-detection

[4] https://github.com/karolpiczak/ESC-50

[5] https://github.com/karolpiczak/ESC-50#download

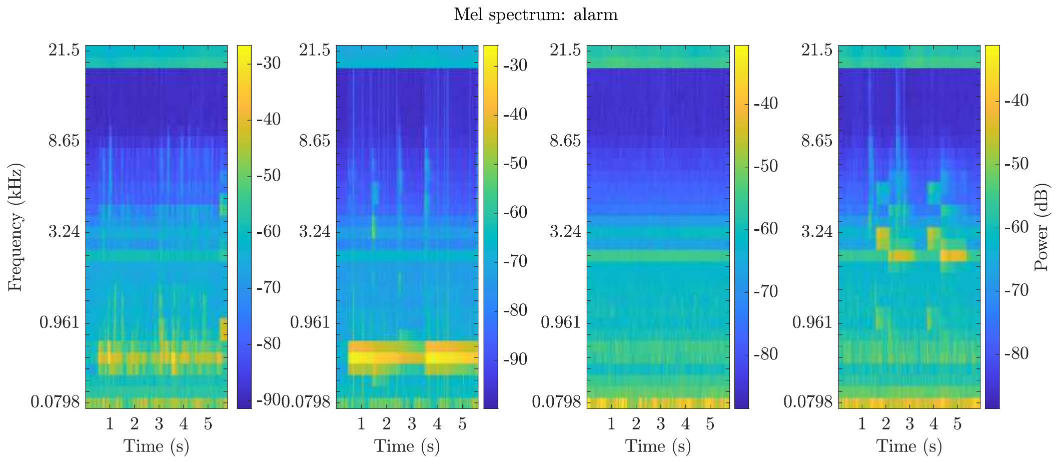


Figure 9.1 Mel spectrum for *alarm*.

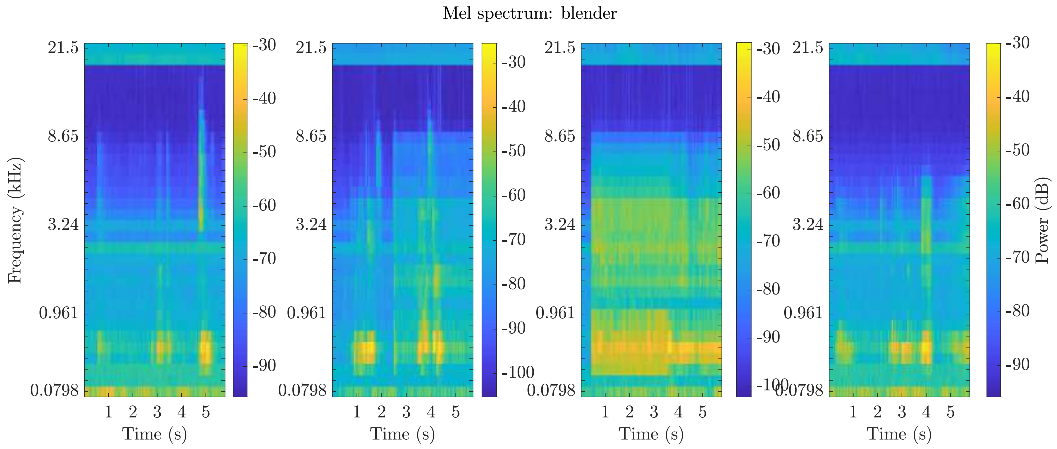


Figure 9.2 Mel spectrum for *blender*.

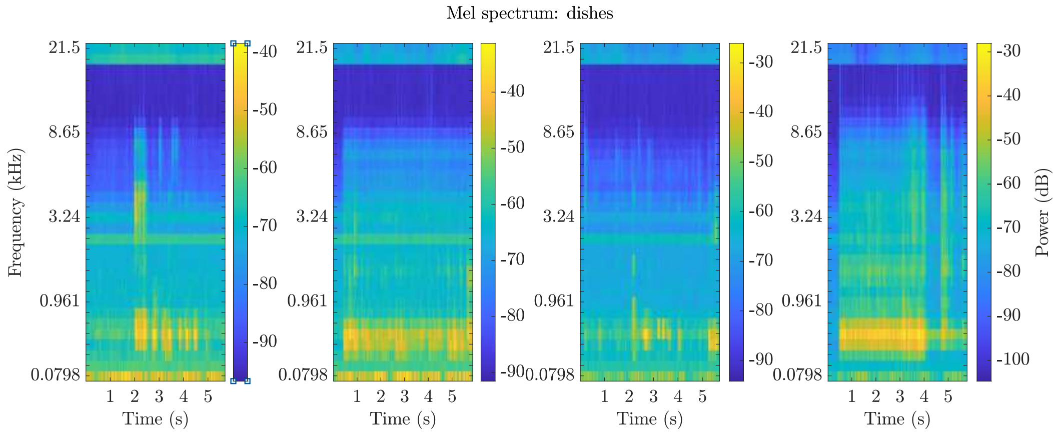


Figure 9.3 Mel spectrum for *dishes*.

Table 9.1 Summary Datasets

	# samples	# classes	length	train	test	sampling rate
dcase_icub	360	4	variable	288	72	48 kHz
ESC-10	400	10	fixed	320	80	44.1 kHz
ESC-50	2000	50	fixed	1600	400	44.1 kHz

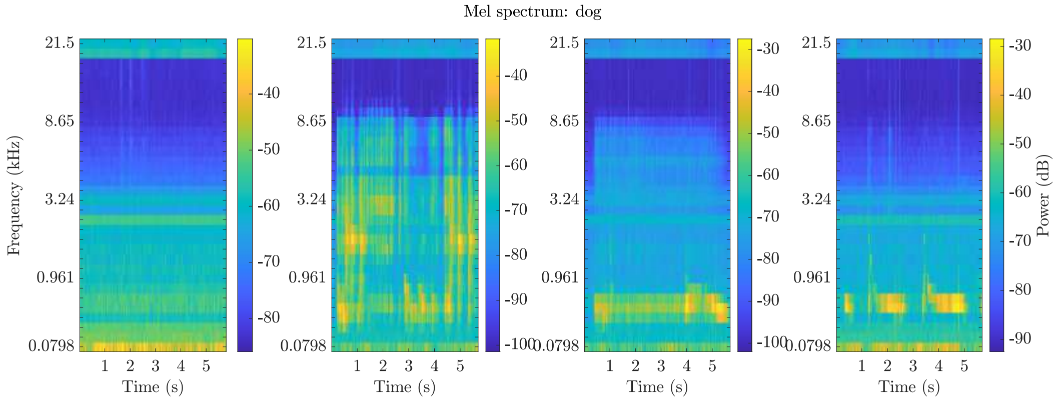


Figure 9.4 Mel spectrum for *dog*.

9.3 Models and Experimental Settings

Different learning strategies have been used. In the following section, a summary of the computational principles of these models will be elaborated. Additionally, the model hyper-parameters, training parameters, and experimental configurations will be reported. Generally, 80% of the data was used for training and 30% for testing. For a fair comparison, the ESC50 data is further split into specific folds following [188]. There are two categories of learning strategies, the first category is classical learning strategies, and the second category is continual learning algorithms.

9.3.1 Learning Models

Matlab 2022a was used to implement the classical learning strategies with the use of echo state network toolbox [96].

k-nearest neighbour knn classifier is an unsupervised learning approach that clusters the feature space to a defined set of clusters and bases the clustering on the distance between the instance and the cluster of the *k* nearest neighbors.

Mel frequency cepstrum coefficients were used as a representation for the input data. For the cross-validation, 5-folds were used. Following [188], the choice for the K values were 10 for the ESC-10 and ESC-50 dataset. For the dcase_icub dataset, 4 folds were used for cross-validation. A preliminary experiment showed that using the Manhattan metric performs superior compared to the Euclidean and squared Euclidean distance. Therefore, it has been used as well as the correlation among data samples.

Echo State Network Echo State Network (ESN) [96] implements a particular training strategy for recurrent neural networks. It has one hidden layer which is called the *reservoir*. The neurons of the hidden layer are randomly initialized in the beginning and remain fixed without changing. When the input is fed to the network, the hidden layer acts as a kernel and transfers the input to the output neurons *readout* as a feature space collected in a design matrix X. Figure 9.5 shows a standard ESN. The training of the ESN follows the equation 9.1.

$$x(t) = f(u(t)W_{in} + \alpha x(t-1)W_{res} + y(t)W_{fb} + \nu(t)) \tag{9.1}$$

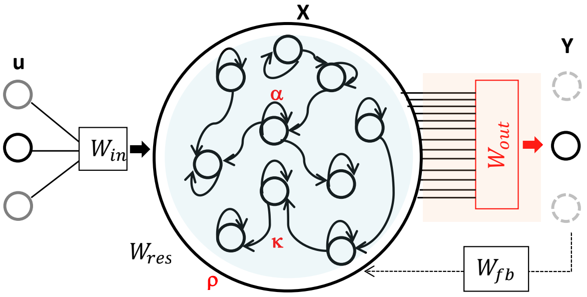


Figure 9.5 Standard echo state network with sparsity hyperparameter κ, leakage α, and spectral radius ρ of reservoir matrix W_{res}. For supervised learning, only the weights W_{out} are trained via regression of the reservoir states X and output Y.

where $f(.)$ is an activation function. *tanh* activation function was used here. and W_* are the layer-wise connectivity matrices for the input, reservoir, and feedback, respectively. $\nu(t)$ is a noise term that had been omitted and not used. Finally, the output is then computed as shown in the equation 9.2.

$$y(n) = g(W_{out}X) \tag{9.2}$$

where g is a defined output function. When g function is linear, and Y (the desired output) is known, training an ESN became a sort of a regression (here regularized [87]):

$$W_{out} = YX^T(XX^T + \lambda \mathbf{I})^{-1} \tag{9.3}$$

where **I** is the identity matrix and T is the transpose operator. For $\lambda = 0$, the regression is calculated using the pseudo-inverse:

$$W_{out} = (YX^{\dagger})^T \tag{9.4}$$

In the experiment, sparsity of $\kappa = 0.1$ was used (10% of the neurons in the hidden layer) and *tanh* activation function. The choice of activation was done based on a preliminary experiment to compare linear activation and *tanh* activation. The results of the preliminary trials showed a better performance using *tanh*. For the spectral radius ρ, and α, a grid search was performed to select suitable values but there wasn't any big difference in the performance, presumably due to the small size dataset. So, the used values were $\rho = 0.9$ and $\alpha = 1$. 20 trials were performed to measure the performance using different hidden layer sizes as the following $N_{res} = \{100, 200, 400, 600, 800\}$.

VGGish network VGGish is a correspondent deep neural network to the popular VGG network used in image classification. The input layer of the VGGish network takes only the coefficients of the Mel spectrum. Therefore, it is important to process the raw input audio to match the input size of the VGGish network. Further, the architecture has six cascaded convolutional layers. Each layer has max-pooling layer and uses reLU activation. After the convolutional stages, the architecture has a fully connected layer. This sum up of nine trainable layers. The output layer is a regression layer that use mean-squared error (MSE) as an optimization minimization goal. In the experiment, the pre-trained VGGish network was used for classification without any modifications. The VGGish network is used here as the upper bound of this study.

9.3.2 Continual Learning Algorithms

Continual learning strategies were developed on python. `Librosa` library [147] was used for the audio preprocessing and the `Avalanche` library [141] for the implementation of the CL strategies as both are python-based code that allows to simplify the experimental pipeline. Opposed to batch supervised learning, CL models are evaluated incrementally by their number of tasks, i.e. the performance metric can be evaluated every time a new class is added. Hence, the number of tasks corresponds to the number of classes for each dataset.

The *incremental classifier and representation learning* (iCaRL) algorithm [202] was the first used CL strategy, which joins regularization and rehearsal strategies. In essence, the algorithm learns continually feature representations jointly trained using an update function at the classification stage which makes iCaRL architecture-agnostic. The trick of adding new classes is that the weight space is separated into a fixed representation of the learned data so far and a weight vector associated with an "exemplar", which is held in memory. In this memory, other classes can be added and the training updated accordingly. Training on the new classes while keeping the information of the previous “old" classes is achieved by a loss function that separates into a typical backpropagation (log) loss term and a knowledge distillation term.

As there are no constraints on the underlying feature extractor $\phi : S \rightarrow \mathbb{R}^d$ for samples S, the algorithm qualifies for a comparison with our techniques using CNN-based audio feature extraction. Specifically for the iCaRL algorithm applied here, we used a multi-layer ResNet with n=5 convolutional layers for the audio feature representation. We set the learning rate $\lambda = 0.1$, the momentum term $m = 0.9$, and weight decay $\delta = 1e-05$ for all datasets and use stochastic gradient descent optimization. The memory size to store the required exemplars is 2000. We ran 50 epochs with a mini-batch size of 16 for the dcase_icub dataset, 70 epochs with a mini-batch size of 128 for the ESC10 dataset, and 70 epochs with a mini-batch size of 32 for the ESC50 dataset over 10 trials each.

Second, GDumb [192] algorithm was also used, which is a very general approach to continual learning as it does not impose constraints on task boundaries, sequence order of the tasks, etc. The approach is memory-based, i.e. some data are sampled in a greedy fashion. Specifically, whenever a new sample from a new class is encountered, the greedy sampler stores up to maximal k samples for that class while removing old samples. Subsequently, a so-called *learner* (basically any neural network) learns a mapping $f_{\theta_t} : x \rightarrow y$, where $x, y \in D_t$ are the samples and labels, respectively, while D_t represents the greedily sampled dataset at time instant t. The prediction $\hat{y}$ for any input x is defined as the Hadamard product $\hat{y} = p \odot m$, where p are the predictions resultant from a softmax layer and $m \in \{0,1\}^{|Y_t|}$ denotes a mask used at inference time. We used similar settings as for the iCaRL trials. The same multi-layer ResNet with n=5 convolutional layers backend was used for feature extraction. We set the learning rate $\lambda = 0.1$ and weight decay $\delta = 1e-05$ for all datasets. I used the *adam* optimizer and ran 10 trials for each dataset each trial consists of 15 epochs with a mini-batch size of 8 for all datasets.

9.4 Results and Evaluation

Table 9.2 reports the averaged performance over the running trials. Also, the guessing (chance) percentage as well as the human accuracy reported for the ESC10 and ESC50. This will be useful to compare the different performances and better understand the results. It is important to note that in ESC50 the human accuracy is far less compared with the ESC10. In the analysis of this Karol Piczak [188] observed that mechanistic sounds produced more misclassification than natural sounds and separated the different classes in the ESC50 into easy, medium, and difficult levels.

As mentioned before, VGGish acts as the upper bound in this experiment. Therefore, it is not a surprise to have the best performance using the VGGish network. However, in dcase_icub and ESC50, the test performance deviated strongly (up to 20 %) when compared to the training performance. This shows that the models don't have a high generalization capability. Conversely, both datasets are small compared to standard machine learning datasets comprising hundreds to thousands of data samples. Additionally, in this study, the goal is not to opt for competitive benchmarking hence with more optimization better results might emerge. Only for ESC-10, the results obtained mirror other results using pre-trained deep learning models.

The unsupervised knn classification strategy and the supervised ESN achieved similar results. For the dcase_icub dataset, the test performance was better using the knn classifier. On the other hand, ESN performed better for the ESC10 dataset. Repeatedly, the test performance was much lower than the training. The low testing performance of ESC50 is aligned with the baseline results reported in [187]. The ESN performance reached 64% for ESC10 which is also aligned with the baseline approaches like the random forest, and support vector machines achieve between 50-70% [187]. For the dcase_icub dataset, from reservoir size res_N : 400 on, the performance doesn't increase even when doubling the reservoir size (more complex model). This result supports the observation of the inter-class variability in the preprocessing stage 9.2 which makes the task difficult to be trained. This is shown in 9.6) comparing the same network with ESC10, increasing the complexity of the model (reservoir size) increases the performance in relatively as shown in 9.7. This is providing more insights for upscaling the model with the ESC50 dataset. It also shows how the number of neurons for the hidden layer (reservoir size) in ESN is a crucial hyperparameter. However, most importantly, the test accuracy does not scale with this observation and it can be anticipated that with increasing reservoir size the network will eventually overfit. Therefore, providing larger reservoirs will not necessarily yield a better feature separation for larger classification

Table 9.2 Summary Table of The Training and Testing Accuracy for All Used Datasets and Strategies

Dataset	dcase		ESC-10		ESC-50	
Chance	25		15		2	
Human	-		95.7		81.3	
Model	train	test	train	test	train	test
kNN	71.25	58.3	78.28	65.00	74.82	29.2
ESN	79.0	57.0	87.0	64.0	-	-
iCaRL	38.78	43.81	63.63	58.00	6.67	10.28
GDumb	74.2	42.8	72.51	33.0	64.34	8.55
VGGish	95.63	76.61	93.13	81.35	88.36	68.12

tasks. Adding to these observations, the other hyperparameters like leakage did not have a huge effect on the performance. Therefore, other models like deepESN [66] or hybrid models incorporating other feature extraction methods will be better suited for datasets like ESC50.

In addition to the evaluation of the classical algorithms in table 9.2, Figure 9.8 shows a comparison between the two CL algorithms. For the iCaRL algorithm, the obtained results are roughly in parallel with the baseline models for the ESC-10 dataset. Although both knn and the ESN perform slightly better, iCaRL does not suffer from the large deviation between train and test performance, hence the algorithm provides better generalization capabilities, one of the desired features in all machine learning applications. Notably, iCaRL operates in an open-set scenario in contrast to the knn and ESN models, which learn in batches and have seen all classes during training. Similarly, GDumb performs inferior to the two baselines knn and ESN, however, it performs way better than the iCaRL algorithm for the dcase_icub. This is an interesting result supporting the authors of GDumb who state that for CL learning a lot of assumptions can be relaxed while still (out)performing other CL algorithms on benchmarks. Also, the superior performance on a high variance dataset such as dcase_icub may show that the sampling strategy is better suited here than for the ESC10, where GDumb clearly performs inferior to iCaRL. However, a large deviation between the train and test results is noted, which means that the algorithm is potentially prone to overfitting. Finally, both continual learning strategies fail in their performance for the ESC50 dataset. While iCaRL does not learn well at all, the behavior of the GDumb algorithm is comparable to the knn classifier, i.e., it achieves good training results but yields mediocre test performance.

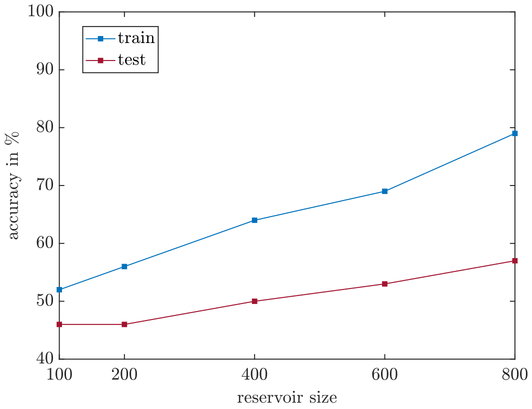


Figure 9.6 Evaluation of the number of reservoir neurons in the baseline ESN on the dcase_icub dataset. While the train performance increases steadily, the test performance stagnates at around 400 neurons, i.e. the reservoir size does not scale with performance.

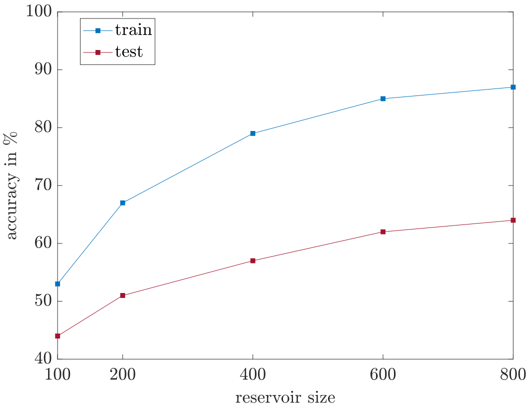


Figure 9.7 Example graph from a trial: Baseline ESN evaluation on the ESC-10 dataset. While the training performance achieves good performance up to $\geq 80\%$, the test performance does not catch up and increases only for a few percentages even when doubling the reservoir size.

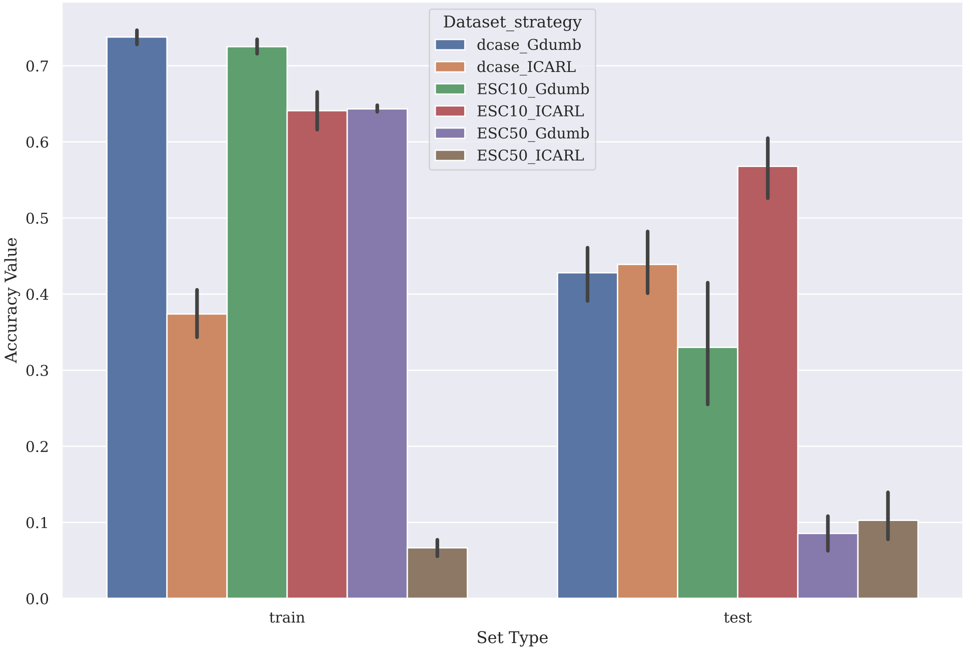


Figure 9.8 Performance comparison for iCaRL and GDumb on all datasets (black bars denote standard deviation). Both algorithms achieve reasonable results comparable to the baseline models knn and ESN, which highlights the potential of continual learning for scenarios with open-set classification as in developmental robotic learning. However, we also observe insufficient performance for the ESC50 dataset, which suggests improvements on the audio feature extraction and loss functions that balance the knowledge about old and new classes.

9.5 Discussion and Future Work

I investigated some shallow models and continual learning algorithms in the context of environmental sound classification with the aim of applying these models to robots as an alternative to the deep learning approach. The models are having the advantage of running in real-time. The results of our experiment showed that both knn classifier and the ESN produce comparatively good results for the dcase_icub and the ESC10 dataset (using the standard audio features). Interestingly, both approaches showed a better results for the ESC10 dataset. This is explainable due to the fact that the recorded dataset (dcase_icub) has a high nose and high variance compared to the clean audio in the ESC10. Both knn and ESN can be considered “lightweight" as their training and test times are low (in the range of ms) compared to intensive training involving GPUs for deep learning applications. But even if the inference time using deep models is reasonable, i.e. applicable in real-time, often the weights used for testing are space-consuming, which renders the direct application on robotic platforms with limited storage capacity difficult. Therefore, these results show a promising direction to further explore those lightweight approaches in the robotic audio domain.

On the other hand, the results showed lower performance using the incremental class learning approach. This means that the model learns first class 1, then class 2, and so on. For the ESC10 dataset, the differences between CL and the baseline methods are relatively small. However, we observed that in CL the model is biased towards some classes more than others during learning which means that representations from subsequent classes are learned insufficiently. This might be due to the data itself, as some classes might be difficult to be recognized in comparison with others. This has been actually pointed out in the dataset characteristics [188]. or example, a *frog* makes a unique, natural sound while a *blender* is mechanistic and can thus be more easily confused with another technical machine.

For both CL strategies (iCaRL and GDumb), the performance for the dcase_icub was lower compared to ESC50. As mentioned before, the main challenge was the high variance in the recorded dataset which negatively impacted the performance. Analysing the strategies individually, the GDumb tends to overfit which was observed by the big gap between the train and test results. This specific result is actually an interesting finding as the authors introducing GDumb state that relaxing some of the assumptions in CL training (e.g. sequence order) do not change performance.

All the methods except VGGish were not successful when the ESC50 dataset was applied. The performance was better than the random guessing (e.g., 29.2% vs. 2% chance for knn), which shows a training trend but it is clearly shown in the results that the baseline approaches and the CL algorithms do not scale up with a large number of classes. Continual learning has been successfully applied to image classification tasks including CIFAR-100 comprising 100 classes but less is known for large-scale applications for audio data. Until now, more processing steps need to be incorporated for larger datasets as also shown by [61], who highlighted that superior performance for ESC50 is achieved by the L3 and SoundNet architectures or using pre-training [85]. The reliable feature representation is still an open question in audio engineering as mentioned by [236]. Using data augmentation [180] methods can also boost performance and we believe that at least the augmentation, possibly coupled with self-supervised learning, is the next important future step building upon our study. Further, more in deep analysis on the regularization and the effect of memory buffer may give us more understanding of the incremental behavior of learning knowledge while retaining the "old" knowledge. In a similar way, ESNs are not capable to be applied to large datasets as we show that when upscaling the number of classes the performance drops. Therefore, we suggest applying architectural variants that are likely able to classify larger audio corpora like deepESN [66].

Our results don't have a competition with the state-of-the-art deep learning results. However, our study shows alternative directions (ESN models and CL algorithms). These directions are computational schemes for robot audition that need a real-time response without the use of high computational power (GPUs). The trade-off here can be acceptable. Our findings have an important role in the direction of developmental robotics research. they advance further research directions into developmental learning in robots, where reservoir-based approaches like ESN and continual learning are underrepresented [100, 132], especially in the audio domain.

Further, we would like to apply continual learning algorithms in a real scenario (on the robot) and also embed them in our architecture. As we have shown for the dcase_icub data, audio models need to capture high signal variances due to possible noise corruptions and the introduction of high-frequency components induced by the robotic hardware (e.g. the fan). Even for a small-scale classification task, we show how these issues negatively the performance even for the VGGish network. Therefore, we see the potential to improve the audio preprocessing pipeline by, e.g., adding denoising autoencoders, which are fast to train in an unsupervised fashion and thus simple to integrate. Concurrently to the practical application of CL, we will analyze further the relaxation of CL assumptions [192] to develop novel strategies to facilitate their deployment in robotic audio scenarios.

9.6 Conclusion

Deep learning models are considered state-of-the-art in many technologies including auditory and visual solutions. However, they do not fit well with human-robot interaction. This is mainly due to the limited availability of data, the computational capacity of robotics, and also unlabeled data. This chapter explained the research conducted to tackle these drawbacks by exploring alternative learning methodologies for audio classification in robotics, specifically shallow models and continuous learning. In this research, we used three different datasets (one is recorded from the microphones of the robot, and two benchmark datasets). The results suggested that the proposed methodologies might have the potential of reaching good performances. The explored methods are less complex and easier to train. Additionally, this research directly impacts the developmental robotic applications that use open-set audio learning. Further development is necessary for big datasets.

Part V

Conclusion

Chapter 10

Summary and Discussion

A hard beginning maketh a good ending.

John Heywood

10.1 Thesis Overview

Robotics are now integrated into human life in various applications. This includes manufacturing, healthcare, education, transportation, and many other domains. In recent years, technological enhancements took the mechanical body and control systems to the required sufficient levels. The real challenge in robotics is now mainly in the cognitive development of robotic agents. For the applications that engage interactions with human partners, the challenge even becomes more complex because it adds the social cognitive components to the requirement.

One of the essential aspects of social interactions is shared perception. Shared perception is the process of understanding a shared context by processing both the environmental cues and the social partner/s states. This process mutually affects the perception of each of the observers. Additionally, it has a direct impact on successful collaborations and teamwork.

In the robotics field, shared perception is often studied from the human side (How the robot changes the perception of the human), or some related skills to share perception (Perspective Taking, gaze understanding, and gaze following). However, endowing the robot with shared perception skills would be the gate to effective collaborations between the robot and the human.

Firstly, We drew general five main skill categories that are required for the robots to endow the robot with perception capabilities that can be shared with social partners. These skills are indeed interconnected and very general. They are listed as follows:

- Having a common representation
- Expressing Effective Communication
- Spatiotemporal Coordination
- Affective modulation mechanism
- Understanding the other

From the cognitive point of view, these skills are in different levels of abstraction and complexity. Additionally, some of them are dependent on each other.

Following the formation of the overall picture of the required skills, my research focuses on building cognitive architectures that are related to the first three skill categories. This selection of these three categories is due to the approach we decided on and the level of complexity the required skill. The approach we decided in has Our approach has four main characteristics **Biological Inspiration.**, **Audio Visual (Multimodality)**, **Generalized Settings**, and **Attention Based**. The biologically inspired model gives the advantage of having a human-like approach which might lead to a better understanding of the biological process and also better socially interactive scenarios. The multi-modality in our approach is targeted as it is a well-known challenge in cognitive architecture and different robotic applications. The same motive is also applied to the generalization point. Finally, we started with attention and based it on it as it is one of the first steps in perception.

Following this approach, I formed the research questions to target different skills within the first three defined categories. The choice of these skills was purely based on their cognitive complexity and level of abstraction as well as our research approach. These skills are as **Having a common representation** which is important to automatically achieve an understanding of the human partner that leads to effective collaboration, **Expressing Effective Communication** which allows the partner to interpret the behavior of the robot correctly, and **Spatiotemporal Coordination** which ensures the coordination in time and space while interacting with the environment which is a crucial aspect in collaboration. The addressed skills in each category are as the following:

- Having a common representation:

    - Common Attentional Representation
    - Common Environmental Representation
    - Common Proactive Behavior

- Expressing Effective Communication:
    - Effective Gaze and Pointing
- Spatiotemporal Coordination:
    - Coordination Attentional Timing
    - High-level Action Coordination

These skills were framed in three research questions. In trying to answer these research questions, I have developed different cognitive architectures that are based on state-of-the-art audio and visual attention models. The development of the architecture is generally in two stages. The first stage targeted audio-visual joint attention and the second stage targeted building an active generic architecture for human-like audio-visual perception. During the development of the architectures, auditory processing was the main bottleneck of the process in terms of time and accuracy. Therefore, I did some research work on how to tackle this issue and proposed alternative development pathways. In the following part, I will draw a conclusion about each of the stages and what are the main conceptual outcomes.

10.1.1 Audio-Visual Joint Attention For Human-Robot Interaction

The first stage of the thesis answered the first research question *RQ1: How can we integrate state-of-the-art vision and audio models to allow the robot to jointly attend to the environment with a human partner? Is the behavior of the robot effectively received by the human partner? and What is the mutual influence between the robot and the human partner during the interaction using this architecture?* . I developed a memory-based audio-visual integrated attention architecture that does have the same attentional characteristics as humans. Further, I tested the mutual influence during an attentional task jointly between the robot and a human partner. The gaze and pointing behavior were effectively perceived by the human partner. Additionally, the robot's behavior influenced the gaze of the human partner.

The response of the robot using the proposed architecture was fairly comparable to the visual stimulus. But much slower than the human when responding to auditory-only stimulation. Additionally, the accuracy of the robot's decision was affected negatively by spatial congruency between the vision and audio. In the second stage of the thesis, the focus was on the robot's perception perspective and improving its integration capabilities.

10.1.2 Active Generic Architecture of Audio-visual Perception

In this stage, another biologically inspired multi-sensory perception architecture was proposed, build, and tested. The architecture is based on the attention architecture proposed in the first stage of the thesis but modifies the multi-sensory integration procedure, and adds more components to the architecture. This version of the architecture adopts the dual pathways (what and where) for both visual and auditory streams, a multi-sensory bayesian hierarchical inference process that takes into consideration the prior perceived experiences. The priors are stored in different kinds of short-term and long-term memory to serve the inference and the association processing. Further, a sub-part of the architecture was used to examine the active behavior of the robot to improve perception or more specifically audio localization. This was to show how common proactive behaviours can be implemented using the proposed architecture. Further, I applied the architecture in two other robots rather than iCub which are Pepper, and the Beast. Additionally, we applied the architecture in a real-world scenario to do multisensory coordination between the Beast and a drone in Tiptree (a vertical strawberry farm in the UK). The applied architecture in the real world closed the full circle between high-level action planning, low-level perception, and decision-making processes. This stage targeted answering the second and the third research questions.

RQ2: How can this integrated audio-visual attention architecture be used by the robot to understand a complex audio-visual environment? How can uncertainty be handled? How can the robot actively perceive the environment?

RQ3: Can this perception architecture be generalized to different robots and applied to a complex task that requires coordination with another agent?

10.1.3 Alternative Auditory Development Pathways

The auditory processing was always the bottleneck of the whole architecture due to its complexity and the high noise-to-signal ratio. Therefore, the last part of the thesis was targeting the improvements in the auditory system. These improvements were generally done in two directions. Firstly, an alternative developmental pipeline for sound source localization. Secondly, alternative approaches for auditory classification (rather than Deep learning models) that are suitable for robotics environments. We suggest in the alternative pipeline the Head-Related Transfer functions (HRTFs) of the robot to train different models. The pipeline goes by firstly measuring the HRTFs of the robot, following this by rendering spatial auditory datasets, and finally training different machine learning models to the required task. We applied this pipeline to create a model for single-speaker localization and exceeded the state-of-the-art performance. This pipeline is tackling different challenges in robot audition. The second part of this stage was exploring different shallow models and continuous learning for environmental sound classification. The motive of this work was due to the current deep learning models that are commonly used in audio processing) do not really fit the robotics application because of the high processing power and size of the model. So, I explored different techniques to develop auditory classification technologies that fit human-robot interaction scenarios.

10.2 The Novelty and Contribution

This research is framing the cognitive architecture development for shared perception in robotics. The contribution of this research is can be framed as the following:

- **The theoretical developmental requirements for human-robot shared perception**. This point has been illustrated in the introduction part of the thesis. The requirements are the five skill categories that the robot requires to enable shared perception mutually between the robot and the human.

- **The memory-based time variant decision-making strategy for audio-visual attention** Although there has been audio-visual architecture for attention, our architecture has a novel biologically inspired decision-making strategy for attentional decision-making.

- **The experimental evaluation of the audio-visual cognitive architecture** We tried to address the mutual influence between the robot (while using the attention architecture) and the human participants in an attentional task that they are jointly doing. This has not been addressed before for the existing architectures.

- **The generic proactive audio-visual architecture for perception**. We designed, implemented, and applied on three different robots and a real-world use case an audio-visual perception cognitive architecture. Unlike the existing architectures and systems, this architecture is not task-specified. It can be applied to various tasks and applications. The architecture also addressed cross-modal interaction which has been ignored in most of the existing architectures. It is also addressing the integration of the priors in different perceptual levels, the proactive behavior to improve perception, and the high-level task coordination and execution. Finally, The implementation of the architecture was designed to be modular and scalable for future development towards the ultimate goal (human-robot shared perception).

- **Alternative auditory pathways** I did research on how to solve current challenges in robot audition by proposing two novel pipelines. The first one is for localization and the second one is for classification, which are the dual pathways of audio in the architecture. The first pipeline will directly help the iCub community in deploying new auditory solutions in their groups. With the second audio developmental direction, I scratched the surface of trying different modalities and showed that alternative methodologies need to be discovered more as they might be sufficient in robotic applications.

10.3 Future Work

In this section, we suggest future work and the use of our research. There are three possible directions of development for this research as the following:

- Integrating other cognitive components and skills. We focused on some skills within three categories out of five required categories in shared perception. Integrating other cognitive components to facilitate other skills will take the robot one more step further for effective collaborations. Perspective-taking, gaze understanding, ad gaze estimation are some skills that are already been studied. So they might be a good start to see how they can be integrated into the architecture.
- Improving the individual modalities. In the architecture, we used simplified models. For example, we used color segmentation as what pathway in vision. However, more complex models can be used to extract more information such as shapes or object recognition, or even more abstract representations for the objects as embedding or other features. This goes the same for the memory and audio. Another possibility at the sensing level is using the latest technologies in sensing the environment through neuromorphic sensing.

- Simulating human perception and impairment. One useful application and area of future development is simulating human perception with the aim of testing cognitive theories and hypotheses or developing new educational technological tools. Additionally, using the architecture to simulate impairments can help in understanding and then designing effective prevention, treatment, and adaptation technologies.

10.4 Publications

- **Workshop Paper** Eldardeer, O., G. Sandini, and F. Rea. "A Biological Inspired Cognitive Model of Multi-sensory Joint Attention in Human-Robot Collaborative Tasks, AVHRC workshop, ROMAN 2020"

- **Workshop Paper** Eldardeer, Omar, et al. "Auditory Perception for Interactive Robots: a Cognitive Framework to Include Motor Commands and Working Memory in the Process of Auditory Sound Localization, Sound in HRI workshop, HRI 2021"

- **Journal Paper** Eldardeer, Omar, et al. "A Biological Inspired Cognitive Framework for Memory-Based Multi-Sensory Joint Attention in Human-Robot Interactive Tasks." Frontiers in Neurorobotics (2021): 133.

- **Submitted Manuscript** Eldardeer et al. "When Deep is not Enough: Understanding of Shallow and Continual Learning Models in Realistic Environmental Sound Classification for Robots" International Journal of Humanoid Robotics.

- **To be submitted (Targeting Journal Paper)** From attention to long-term memory, a general multimodal cognitive architecture for robotic life-long perception.

- **To be submitted (Targeting Journal Paper)** Biologically Inspired Auditory-Visual Perception framework for Goal-directed coordination between drones and mobile robots in Smart Farms

- **To be submitted (Targeting Conference Paper)** Noise robust model for multiple speakers' simultaneous identification and localization for robots

- **To be submitted (Targeting Conference Paper)** Towards A Human-Like Shared Perception in Robotics: a Developmental Road Map

Chapter 11

Epilogue

لا يستطاع العلم براحة الجسم
Knowledge can not be attained with bodily comfort

ابن كثير
Ibn Kathir

Reaching this point in the thesis, firstly, I would like to express my gratitude for your time and efforts in reading my thesis. Then I seize this opportunity to share my thoughts, lessons, and experiences throughout my journey.

Indeed the journey was not easy. There were many challenges in different dimensions to overcome. Socially being away from my family and friends and doing this research work in a country that holds a different culture, language, and individual interests, scientifically coming from a different background that is far from cognitive studies, and mentally, in the period of COVID-19.

Science has an effective value in my personal beliefs. It is a shared language that brings individuals from all over the world together. This belief is what drives my passion for scientific research, as I feel responsible for doing my best to make a positive impact on society and individuals worldwide. I firmly believe that science is not just an academic pursuit, but a tool for building peaceful and inclusive communities.

Having this thought in my mind was the main supporting element to overcome all the challenges. This is also the reason that I am committed to volunteering and giving back to society. I am currently serving as the public relations team leader for Egypt Scholar Inc., an NGO that supports scientific communities in Egypt and the Arabic world. This volunteering experience has allowed me to further develop my leadership skills while contributing to a cause that I am deeply passionate about. Ultimately, my goal as a scientist and a volunteer is to create a world in which everyone has the opportunity to thrive and succeed without any discrimination.

Family, friends, and colleagues provided me with invaluable sources of support and encouragement. They offered a listening ear, a backbone to stand on, inspiration, and practical assistance when needed. I am always grateful to all of them, words are incapable of thanking them enough. I feel truly blessed to have you all in my life. THANK YOU ALL.

I hope reading this thesis was a pleasant experience for you. I also wish that this thesis contributes to creating a positive impact in the field, inspires other researchers and students, and be my first step toward better achievements.

و الحمد لله رب العالمين

References

[1] Admoni, H. and Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63.

[2] Ajoudani, A., Zanchettin, A. M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., and Khatib, O. (2018). Progress and prospects of the human-robot collaboration. *Autonomous Robots*, 42:957–975.

[3] Al-Azzawi, N., Bayram, B., and Ince, G. (2018). Audiovisual attention for robots from a developmental perspective. In *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, pages 312–317. IEEE.

[4] Alais, D. and Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3):257–262.

[5] Ali, Z. and Talha, M. (2018). Innovative method for unsupervised voice activity detection and classification of audio segments. *Ieee Access*, 6:15494–15504.

[6] Anderson, J. R., Matessa, M., and Lebiere, C. (1997). Act-r: A theory of higher level cognition and its relation to visual attention. *Human–Computer Interaction*, 12(4):439–462.

[7] Anzalone, S. M., Ivaldi, S., Sigaud, O., and Chetouani, M. (2013). Multimodal people engagement with icub. In *Biologically Inspired Cognitive Architectures 2012: Proceedings of the Third Annual Meeting of the BICA Society*, pages 59–64. Springer.

[8] Argentieri, S., Danes, P., and Soueres, P. (2015). A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech and Language*, 34:87 – 112.

[9] Avillac, M., Olivier, E., Denève, S., Ben Hamed, S., and Duhamel, J.-R. (2004). Multisensory integration in multiple reference frames in the posterior parietal cortex. *Cognitive Processing*, 5:159–166.

[10] Awh, E. and Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in cognitive sciences*, 5(3):119–126.

[11] Awh, E., Vogel, E. K., and Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience*, 139(1):201–208.

[12] Bach, P. and Schenke, K. C. (2017). Predictive social perception: Towards a unifying framework from action observation to person knowledge. *Social and Personality Psychology Compass*, 11(7):e12312.

[13] Baldassarre, G., Lord, W., Granato, G., and Santucci, V. G. (2019). An embodied agent learning affordances with intrinsic motivations and solving extrinsic tasks with attention and one-step planning. *Frontiers in Neurorobotics*, 13:45.

[14] Balint, T. and Allbeck, J. M. (2013). What's going on? multi-sense attention for virtual agents. In *Intelligent Virtual Agents: 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013. Proceedings 13*, pages 349–357. Springer.

[15] Batteau, D. W. (1967). The role of the pinna in human localization. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 168(1011):158–180.

[16] Battich, L., Fairhurst, M., and Deroy, O. (2020). Coordinating attention requires coordinated senses. *Psychonomic bulletin & review*, 27(6):1126–1138.

[17] Baumann, C., Rogers, C., and Massen, F. (2015). Dynamic binaural sound localization based on variations of interaural time delays and system rotations. *Journal of the Acoustical Society of America*, 138(2):635–650.

[18] Bhat, A. A. and Mohan, V. (2018). Goal-directed reasoning and cooperation in robots in shared workspaces: an internal simulation based neural framework. *Cognitive computation*, 10(4):558–576.

[19] Biddulph, A., Houliston, T., Mendes, A., and Chalup, S. K. (2018). Comparing computing platforms for deep learning on a humanoid robot. In *International Conference on Neural Information Processing*, pages 120–131. Springer.

[20] Bigelow, A. E., MacLean, K., and Proctor, J. (2004). The role of joint attention in the development of infants' play with objects.

[21] Blauert, E. (1997a). *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press.

[22] Blauert, J. (1997b). *Spatial hearing: the psychophysics of human sound localization*. MIT press.

[23] Bodini, M. (2019). Sound classification and localization in service robots with attention mechanisms. In *Computer-Aided Developments: Electronics and Communication*, pages 68–76. CRC Press.

[24] Bonarini, A. (2020). Communication in human-robot interaction. *Current Robotics Reports*, 1(4):279–285.

[25] Breazeal, C., Edsinger, A., Fitzpatrick, P., and Scassellati, B. (2001). Active vision for sociable robots. *IEEE Transactions on systems, man, and cybernetics-part A: Systems and Humans*, 31(5):443–453.

[26] Breazeal, C. and Scassellati, B. (1999). A context-dependent attention system for a social robot. *rn*, 255(3).

[27] Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B., and Zelinsky, G. J. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition*, 106(3):1465–1477.

[28] Browne, W., Kawamura, K., Krichmar, J., Harwin, W., and Wagatsuma, H. (2009). Cognitive robotics: new insights into robot and human intelligence by reverse engineering brain functions [from the guest editors]. *IEEE Robotics & Automation Magazine*, 16(3):17–18.

[29] Bruinsma, Y., Koegel, R. L., and Koegel, L. K. (2004). Joint attention and children with autism: A review of the literature. *Mental retardation and developmental disabilities research reviews*, 10(3):169–175.

[30] Butcher, A., Govenlock, S. W., and Tata, M. S. (2011). A lateralized auditory evoked potential elicited when auditory objects are defined by spatial motion. *Hearing research*, 272(1-2):58–68.

[31] Butterworth, G. (1991). The ontogeny and phylogeny of joint visual attention.

[32] Cangelosi, A. and Asada, M. (2022). *Cognitive robotics*. MIT Press.

[33] Cech, J., Mittal, R., Deleforge, A., Sanchez-Riera, J., Alameda-Pineda, X., and Horaud, R. (2013). Active-speaker detection and localization with microphones and cameras embedded into a robotic head. In *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 203–210. IEEE.

[34] Chambers, C., Sokhey, T., Gaebler-Spira, D., and Kording, K. P. (2018). The development of bayesian integration in sensorimotor estimation. *Journal of Vision*, 18(12):8–8.

[35] Chen, Y.-C., Huang, S.-F., Lee, H.-y., Wang, Y.-H., and Shen, C.-H. (2019). Audio word2vec: Sequence-to-sequence autoencoding for unsupervised learning of audio segmentation and representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(9):1481–1493.

[36] Chen, Y.-C. and Spence, C. (2017). Assessing the role of the 'unity assumption'on multisensory integration: A review. *Frontiers in psychology*, 8:445.

[37] Chevalier, P., Kompatsiari, K., Ciardo, F., and Wykowska, A. (2020). Examining joint attention with the use of humanoid robots-a new approach to study fundamental mechanisms of social cognition. *Psychonomic Bulletin & Review*, 27(2):217–236.

[38] Chong, H.-Q., Tan, A.-H., and Ng, G.-W. (2007). Integrated cognitive architectures: a survey. *Artificial Intelligence Review*, 28:103–130.

[39] Churchland, A. K., Kiani, R., and Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, 11(6):693–702.

[40] Colby, C. L. (1998). Action-oriented spatial reference frames in cortex. *Neuron*, 20(1):15–24.

[41] Connor, S. (2000). *Dumbstruck-A Cultural history of ventriloquism*. OUP Oxford.

[42] Constantinidis, C. and Wang, X.-J. (2004). A neural circuit basis for spatial working memory. *The Neuroscientist*, 10(6):553–565.

[43] Cook, M. P. (2006). Visual representations in science education: The influence of prior knowledge and cognitive load theory on instructional design principles. *Science Education*, 90(6):1073–1091.

[44] De Giacomo, G. (1998). Cognitive robotics. In *Proc. AAAI'98 Fall Symposium, Technical Report FS-98-02, AAAI Press, Menlo Park, California.*

[45] De Lange, F. P., Heilbron, M., and Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, 22(9):764–779.

[46] Deák, G. O., Fasel, I., and Movellan, J. (2001). The emergence of shared attention: Using robots to test developmental theories. In *Proceedings 1st International Workshop on Epigenetic Robotics: Lund University Cognitive Studies*, volume 85, pages 95–104.

[47] Delange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[48] Deneve, S. and Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology-Paris*, 98(1-3):249–258.

[49] Ditterich, J. (2006). Evidence for time-variant decision making. *European Journal of Neuroscience*, 24(12):3628–3641.

[50] Driver, J. and Spence, C. (1998). Cross–modal links in spatial attention. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1373):1319–1331.

[51] Duhamel, J.-R., Bremmer, F., Ben Hamed, S., and Graf, W. (1997). Spatial invariance of visual receptive fields in parietal cortex neurons. *Nature*, 389(6653):845–848.

[52] Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of vision*, 7(5):7–7.

[53] Ernst, M. O. and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433.

[54] Eskimez, S. E., Duan, Z., and Heinzelman, W. (2018). Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5099–5103. IEEE.

[55] Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108*. Audio Engineering Society.

[56] Ferrone, L. and Zanzotto, F. M. (2020). Symbolic, distributed, and distributional representations for natural language processing in the era of deep learning: A survey. *Frontiers in Robotics and AI*, 6:153.

[57] Fetsch, C. R., Pouget, A., DeAngelis, G. C., and Angelaki, D. E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature neuroscience*, 15(1):146–154.

[58] Fischer, T. and Demiris, Y. (2019). Computational modeling of embodied visual perspective taking. *IEEE Transactions on Cognitive and Developmental Systems*, 12(4):723–732.

[59] Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166.

[60] Fougnie, D. (2008). The relationship between attention and working memory. *New research on short-term memory*, 1:45.

[61] Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., and Schuller, B. (2017). audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, 18(1):6340–6344.

[62] French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

[63] Frischen, A., Bayliss, A. P., and Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological bulletin*, 133(4):694.

[64] Fu, D., Weber, C., Yang, G., Kerzel, M., Nan, W., Barros, P., Wu, H., Liu, X., and Wermter, S. (2020). What can computational models learn from human selective attention? a review from an audiovisual unimodal and crossmodal perspective. *Frontiers in Integrative Neuroscience*, 14:10.

[65] Gaggioli, A., Chirico, A., Di Lernia, D., Maggioni, M. A., Malighetti, C., Manzi, F., Marchetti, A., Massaro, D., Rea, F., Rossignoli, D., et al. (2021). Machines like us and people like you: Toward human-robot shared experience.

[66] Gallicchio, C., Micheli, A., and Pedrelli, L. (2018). Design of deep echo state networks. *Neural Networks*, 108:33–47.

[67] Gan, C., Zhang, Y., Wu, J., Gong, B., and Tenenbaum, J. B. (2020). Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707. IEEE.

[68] Gau, R. and Noppeney, U. (2016). How prior expectations shape multisensory perception. *Neuroimage*, 124:876–886.

[69] Goertzel, B., Lian, R., Arel, I., De Garis, H., and Chen, S. (2010). A world survey of artificial brain projects, part ii: Biologically inspired cognitive architectures. *Neurocomputing*, 74(1-3):30–49.

[70] Gonzalez-Billandon, J., Belgiovine, G., Tata, M., Sciutti, A., Sandini, G., and Rea, F. (2021). Self-supervised learning framework for speaker localisation with a humanoid robot. In *2021 IEEE International Conference on Development and Learning (ICDL)*, pages 1–7.

[71] Gonzalez-Billandon, J., Grasse, L., Sciutti, A., Tata, M., and Rea, F. (2019). Cognitive architecture for joint attentional learning of word-object mapping with a humanoid robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, Workshop on Deep Probabilistic Generative Models for Cognitive Architecture in Robotics*.

[72] Gonzalez-Billandon, J., Sciutti, A., Tata, M., Sandini, G., and Rea, F. (2020). Audiovisual cognitive architecture for autonomous learning of face localisation by a humanoid robot. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5979–5985. IEEE.

[73] Gori, M., Sciutti, A., Burr, D., and Sandini, G. (2011). Direct and indirect haptic calibration of visual size judgments. *PLoS One*, 6(10):e25599.

[74] Griffiths, T. D. and Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, 5(11):887–892.

[75] Grothe, B., Pecka, M., and McAlpine, D. (2010). Mechanisms of sound localization in mammals. *Physiological Reviews*, 90:983 – 1012.

[76] Groves, P. M. and Thompson, R. F. (1970). Habituation: a dual-process theory. *Psychological review*, 77(5):419.

[77] Grumiaux, P.-A., Kitić, S., Girin, L., and Guérin, A. (2022). A survey of sound source localization with deep learning methods. *The Journal of the Acoustical Society of America*, 152(1):107–151.

[78] Guzhov, A., Raue, F., Hees, J., and Dengel, A. (2021). Esresnet: Environmental sound classification based on visual domain models. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4933–4940. IEEE.

[79] Hambrook, D. A., Ilievski, M., Mosadeghzad, M., and Tata, M. S. (2017). A bayesian computational basis for auditory selective attention using head rotation and the interaural time-difference cue. *PLoS One*, 12(10).

[80] Hartland, C. and Bredeche, N. (2007). Using echo state networks for robot navigation behavior acquisition. In *2007 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 201–206. IEEE.

[81] Haykin, S. and Chen, Z. (2005). The cocktail party problem. *Neural computation*, 17(9):1875–1902.

[82] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[83] He, W., Motlicek, P., and Odobez, J.-M. (2018). Deep neural networks for multiple speaker detection and localization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 74–79. IEEE.

[84] Herr, D., Godel, M., Perkins, R., Pate, L., and Hall, T. (2020). The economic impact of robotics & autonomous systems across uk sectors.

[85] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

[86] Hinaut, X., Petit, M., Pointeau, G., and Dominey, P. F. (2014). Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks. *Frontiers in neurorobotics*, 8:16.

[87] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[88] Hornstein, J., Lopes, M., Santos-Victor, J., and Lacerda, F. (2006). Sound localization for humanoid robots - building audio-motor maps based on the hrtf. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1170–1176.

[89] Hosangadi, R. (2019). A proposed method for acoustic source localization in search and rescue robot. In *Proceedings of the 5th International Conference on Mechatronics and Robotics Engineering*, pages 134–140.

[90] Hrncir, M., Barth, F. G., and Tautz, J. (2005). 32 vibratory and airborne-sound signals in bee communication (hymenoptera). *Insect sounds and communication: physiology, behaviour, ecology, and evolution*, page 421.

[91] Hudson, M., Nicholson, T., Ellis, R., and Bach, P. (2016). I see what you say: Prior knowledge of other's goals automatically biases the perception of their actions. *Cognition*, 146:245–250.

[92] Hunt, J. and Richard, F.-J. (2013). Intracolony vibroacoustic communication in social insects. *Insectes Sociaux*, 60:403–417.

[93] Imai, M., Ono, T., and Ishiguro, H. (2003). Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transactions on Industrial Electronics*, 50(4):636–643.

[94] Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.

[95] Jack, C. E. and Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the “ventriloquism” effect. *Perceptual and Motor Skills*, 37(3):967–979. PMID: 4764534.

[96] Jaeger, H. (2002). Tutorial on training recurrent neural networks, covering bptt, rtrl, ekf and the "echo state network" approach. Technical report, German National Research Center for Information Technology.

[97] Jati, A. and Georgiou, P. (2019). Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1577–1589.

[98] Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*.

[99] Ji, Y., Yang, Y., Shen, F., Shen, H. T., and Li, X. (2019). A survey of human action analysis in hri applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(7):2114–2128.

[100] Jirak, D. and Wermter, S. (2017). Potentials and limitations of deep neural networks for cognitive robots.

[101] Jording, M., Hartz, A., Bente, G., Schulte-Rüther, M., and Vogeley, K. (2018). The "social gaze space": A taxonomy for gaze-based communication in triadic interactions. *Frontiers in psychology*, 9:226.

[102] Kaas, J. H. and Hackett, T. A. (1999). 'what'and'where'processing in auditory cortex. *Nature neuroscience*, 2(12):1045–1047.

[103] Kabir, M. S., Ndukwe, I. K., and Awan, E. Z. S. (2021). Deep learning inspired vision based frameworks for drone detection. In *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pages 1–5. IEEE.

[104] Kaiser, M. K. and Proffitt, D. R. (1984). The development of sensitivity to causally relevant dynamic information. *Child Development*, pages 1614–1624.

[105] Kalinli, O. and Narayanan, S. S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In *INTERSPEECH*, pages 1941–1944.

[106] Kaluarachchi, T., Reis, A., and Nanayakkara, S. (2021). A review of recent deep learning approaches in human-centered machine learning. *Sensors*, 21(7):2514.

[107] Kaplan, F. and Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, 7(2):135–169.

[108] Kaya, E. M. and Elhilali, M. (2017). Modelling auditory attention. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160101.

[109] Keyrouz, F. (2014). Advanced binaural sound localization in 3-d for humanoid robots. *IEEE Transactions on Instrumentation and Measurement*, 63(9):2098–2107.

[110] Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44.

[111] Kim, H.-D., Choi, J.-S., and Kim, M. (2006). Speaker localization among multi-faces in noisy environment by audio-visual integration. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 1305–1310. IEEE.

[112] Kim, U.-H., Nakadai, K., and Okuno, H. G. (2015). Improved sound source localization in horizontal plane for binaural robot audition. *Applied Intelligence*, 42:63–74.

[113] Kirchner, W. (1993). Acoustical communication in honeybees. *Apidologie*, 24(3):297–307.

[114] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

[115] Klein, R. M. (2000). Inhibition of return. *Trends in cognitive sciences*, 4(4):138–147.

[116] Knill, D. C. and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, 27(12):712–719.

[117] Knill, D. C. and Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision research*, 43(24):2539–2558.

[118] Kompatsiari, K., Bossi, F., and Wykowska, A. (2021). Eye contact during joint attention with a humanoid robot modulates oscillatory brain activity. *Social cognitive and affective neuroscience*, 16(4):383–392.

[119] Kothig, A., Ilievski, M., Grasse, L., Rea, F., and Tata, M. (2019). A bayesian system for noise-robust binaural sound localisation for humanoid robots. In *2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, pages 1–7. IEEE.

[120] Kotseruba, I. and Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, 53(1):17–94.

[121] Kozima, H. and Yano, H. (2001). A robot that learns to communicate with human caregivers. In *Proceedings of the First International Workshop on Epigenetic Robotics*, volume 2001.

[122] Krichmar, J. L. (2012). Design principles for biologically inspired cognitive robotics. *Biologically Inspired Cognitive Architectures*, 1:73–81.

[123] Kubovy, M. and Van Valkenburg, D. (2001). Auditory and visual objects. *Cognition*, 80(1-2):97–126.

[124] Kumon, M. and Uozumi, S. (2011). Binaural localization for a mobile sound source. *Journal of Biomechanical Science and Engineering*, 6(1):26 – 39.

[125] Laird, J. E. (2019). *The Soar cognitive architecture*. MIT press.

[126] Lakatos, G., Wood, L. J., Syrdal, D. S., Robins, B., Zaraki, A., and Dautenhahn, K. (2021). Robot-mediated intervention can assist children with autism to develop visual perspective taking skills. *Paladyn, Journal of Behavioral Robotics*, 12(1):87–101.

[127] Langley, P., Laird, J. E., and Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2):141–160.

[128] Lanillos, P., Ferreira, J. F., and Dias, J. (2015). Designing an artificial attention system for social robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4171–4178. IEEE.

[129] Lanillos, P., Meo, C., Pezzato, C., Meera, A. A., Baioumy, M., Ohata, W., Tschantz, A., Millidge, B., Wisse, M., Buckley, C. L., et al. (2021). Active inference in robotics and artificial agents: Survey and challenges. *arXiv preprint arXiv:2112.01871*.

[130] Lastrico, L., Garello, L., Rea, F., Noceti, N., Mastrogiovanni, F., Sciutti, A., and Carfì, A. (2022). Robots with different embodiments can express and influence carefulness in object manipulation. In *2022 IEEE International Conference on Development and Learning (ICDL)*, pages 280–286. IEEE.

[131] Leaver, A. M. and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *Journal of Neuroscience*, 30(22):7604–7612.

[132] Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., and Díaz-Rodríguez, N. (2020). Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58:52–68.

[133] Levesque, H. and Lakemeyer, G. (2008). Cognitive robotics. *Foundations of artificial intelligence*, 3:869–886.

[134] Li, M., Yang, B., Levy, J., Stolcke, A., Rozgic, V., Matsoukas, S., Papayiannis, C., Bone, D., and Wang, C. (2021). Contrastive unsupervised learning for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6329–6333. IEEE.

[135] Li, S. and Peissig, J. (2020). Measurement of head-related transfer functions: A review. *Applied Sciences*, 10(14):5014.

[136] Liaquat, M. U., Munawar, H. S., Rahman, A., Qadir, Z., Kouzani, A. Z., and Mahmud, M. P. (2021). Localization of sound sources: A systematic review. *Energies*, 14(13):3910.

[137] Liepelt, R., Cramon, D., and Brass, M. (2008). What is matched in direct matching? intention attribution modulates motor priming. *Journal of Experimental Psychology: human perception and performance*, 34(3):578.

[138] Lim, A. and Okuno, H. G. (2014). The mei robot: towards using motherese to develop multimodal emotional intelligence. *IEEE Transactions on Autonomous Mental Development*, 6(2):126–138.

[139] Lindauer, M. (2013). Communication among social bees. In *Communication among Social Bees*. Harvard University Press.

[140] Litovsky, R. (2015). Development of the auditory system. *Handbook of clinical neurology*, 129:55–72.

[141] Lomonaco, V., Pellegrini, L., Cossu, A., Carta, A., Graffieti, G., Hayes, T. L., De Lange, M., Masana, M., Pomponi, J., Van de Ven, G. M., et al. (2021). Avalanche: an end-to-end library for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3600–3610.

[142] Lupiáñez, J., Klein, R. M., and Bartolomeo, P. (2006). Inhibition of return: Twenty years after. *Cognitive neuropsychology*, 23(7):1003–1014.

[143] MacDonald, J. and McGurk, H. (1978). Visual influences on speech perception processes. *Perception & psychophysics*, 24(3):253–257.

[144] Matarese, M., Rea, F., and Sciutti, A. (2022). Perception is only real when shared: A mathematical model for collaborative shared perception in human-robot interaction. *Frontiers in Robotics and AI*, 9.

[145] Mazzola, C., Rea, F., and Sciutti, A. (2022). Shared perception is different from individual perception: a new look on context dependency. *IEEE Transactions on Cognitive and Developmental Systems*.

[146] McCallum, M. C. (2019). Unsupervised learning of deep features for music segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 346–350. IEEE.

[147] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.

[148] McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–8.

[149] Meltzoff, A. N. (2007). 'like me': a foundation for social cognition. *Developmental science*, 10(1):126–134.

[150] Metta, G., Fitzpatrick, P., and Natale, L. (2006). Yarp: yet another robot platform. *International Journal of Advanced Robotic Systems*, 3(1):8.

[151] Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., Von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J., et al. (2010). The icub humanoid robot: An open-systems platform for research in cognitive development. *Neural networks*, 23(8-9):1125–1134.

[152] Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). The icub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems*, pages 50–56.

[153] Meyer, M., Beutel, J., and Thiele, L. (2017). Unsupervised feature learning for audio analysis. *arXiv preprint arXiv:1712.03835*.

[154] Miller, G. A., Eugene, G., and Pribram, K. H. (1960). In *Plans and the Structure of Behaviour*. Henry Holt and Co.

[155] Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417.

[156] Mohan, V., Morasso, P., Metta, G., et al. (2011). The distribution of rewards in sensorimotor maps acquired by cognitive robots through exploration. *Neurocomputing*, 74(17):3440–3455.

[157] Moon, A., Troniak, D. M., Gleeson, B., Pan, M. K., Zheng, M., Blumer, B. A., MacLean, K., and Croft, E. A. (2014). Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 334–341.

[158] Moore, C., Dunham, P. J., and Dunham, P. (1995). *Joint attention: Its origins and role in development*. Psychology Press.

[159] Morey, C. C. (2009). Integrated cross-domain object storage in working memory: Evidence from a verbal–spatial memory task. *Quarterly Journal of Experimental Psychology*, 62(11):2235–2251.

[160] Mundy, P. (2018). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder. *European Journal of Neuroscience*, 47(6):497–514.

[161] Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., and Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child development*, 78(3):938–954.

[162] Mundy, P., Card, J., and Fox, N. (2000). Eeg correlates of the development of infant joint attention skills. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 36(4):325–338.

[163] Mundy, P. and Newell, L. (2007). Attention, joint attention, and social cognition. *Current directions in psychological science*, 16(5):269–274.

[164] Mundy, P., Sigman, M., and Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and developmental Disorders*, 20(1):115–128.

[165] Murphy, P. R., Boonstra, E., and Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nature communications*, 7(1):1–15.

[166] Nagai, Y., Hosoda, K., Morita, A., and Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229.

[167] Nakadai, K., Okuno, H. G., and Mizumoto, T. (2017). Development, deployment and applications of robot audition open source software hark. *Journal of Robotics and Mechatronics*, 29(1):16–25.

[168] Nakadai, K., Takahashi, T., Okuno, H. G., Nakajima, H., Hasegawa, Y., and Tsujino, H. (2010). Design and implementation of robot audition system'hark'—open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761.

[169] Nakamura, K., Nakadai, K., Asano, F., and Ince, G. (2011). Intelligent sound source localization and its application to multimodal human tracking. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 143–148. IEEE.

[170] Naranjo-Alcazar, J., Perez-Castanos, S., Zuccarello, P., Antonacci, F., and Cobos, M. (2020). Open set audio classification using autoencoders trained on few data. *Sensors*, 20(13):3741.

[171] Neumann, M. and Vu, N. T. (2019). Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7390–7394. IEEE.

[172] Nguyen, Q., Yun, S.-S., and Choi, J. (2014). Audio-visual integration for human-robot interaction in multi-person scenarios. In *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, pages 1–4. IEEE.

[173] Nikolaidis, S., Zhu, Y. X., Hsu, D., and Srinivasa, S. (2017). Human-robot mutual adaptation in shared autonomy. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pages 294–302.

[174] Nothdurft, H. (1991). Texture segmentation and pop-out from orientation contrast. *Vision Research*, 31(6):1073–1078.

[175] Oberauer, K. (2019). Working memory and attention—a conceptual analysis and review. *Journal of cognition*.

[176] Ognibene, D. and Baldassare, G. (2015). Ecological active vision: Four bioinspired principles to integrate bottom–up and adaptive top–down attention tested with a simple camera-arm robot. *IEEE Transactions on Autonomous Mental Development*, 7(1):3–25.

[177] Okuno, H. G. and Nakadai, K. (2015). Robot audition: Its rise and perspectives. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5610–5614. IEEE.

[178] Orabona, F., Metta, G., and Sandini, G. (2005). Object-based visual attention: a model for a behaving robot. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 89–89. IEEE.

[179] O'Reilly, R. C., Hazy, T. E., and Herd, S. A. (2016). The leabra cognitive architecture: How to play 20 principles with nature. *The Oxford handbook of cognitive science*, 91:91–116.

[180] Palanisamy, K., Singhania, D., and Yao, A. (2020). Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*.

[181] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

[182] Parisi, G. I., Barros, P., Fu, D., Magg, S., Wu, H., Liu, X., and Wermter, S. (2018). A neurorobotic experiment for crossmodal conflict resolution in complex environments. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2330–2335. IEEE.

[183] Pellegrini, L., Graffieti, G., Lomonaco, V., and Maltoni, D. (2020). Latent replay for real-time continual learning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10203–10209. IEEE.

[184] Pelosin, F., Jha, S., Torsello, A., Raducanu, B., and van de Weijer, J. (2022). Towards exemplar-free continual learning in vision transformers: an account of attention, functional and weight regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3820–3829.

[185] Pfeiffer-Leßmann, N., Pfeiffer, T., and Wachsmuth, I. (2012). An operational model of joint attention–timing of the initiate-act in interactions with a virtual human. *Proceedings of KogWis 2012*, page 96.

[186] Phillips, J. L. and Noelle, D. C. (2005). A biologically inspired working memory framework for robots. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 599–604. IEEE.

[187] Piczak, K. J. (2015a). Environmental sound classification with convolutional neural networks. In *2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE.

[188] Piczak, K. J. (2015b). Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018.

[189] Posner, M. I., Cohen, Y., et al. (1984). Components of visual orienting. *Attention and performance X: Control of language processes*, 32:531–556.

[190] Posner, M. I., Rafal, R. D., Choate, L. S., and Vaughan, J. (1985). Inhibition of return: Neural basis and function. *Cognitive neuropsychology*, 2(3):211–228.

[191] Pouget, A., Deneve, S., and Duhamel, J.-R. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nature Reviews Neuroscience*, 3(9):741–747.

[192] Prabhu, A., Torr, P. H., and Dokania, P. K. (2020). Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, pages 524–540. Springer.

[193] Premack, D. and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

[194] Quak, M., London, R. E., and Talsma, D. (2015). A multisensory perspective of working memory. *Frontiers in human neuroscience*, 9:197.

[195] Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A. Y., et al. (2009). Ros: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan.

[196] Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., Coppola, G., Geyer, M. A., Glanzman, D. L., Marsland, S., McSweeney, F. K., Wilson, D. A., Wu, C.-F., and Thompson, R. F. (2009). Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory*, 92(2):135–138. Special Issue: Neurobiology of Habituation.

[197] Rascon, C. and Meza, I. (2017). Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210.

[198] Ratajczak, R., Pellerin, D., Labourey, Q., and Garbay, C. (2016). A fast audiovisual attention model for human detection and localization on a companion robot. In *VISUAL 2016-The First International Conference on Applications and Systems of Visual Paradigms (VISUAL 2016)*.

[199] Ratcliff, R. and Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, 111(2):333.

[200] Rea, F., Kothig, A., Grasse, L., and Tata, M. (2020). Speech envelope dynamics for noise-robust auditory scene analysis in robotics. *International Journal of Humanoid Robotics*, 17(06):2050023.

[201] Rea, F., Sandini, G., and Metta, G. (2014). Motor biases in visual attention for a humanoid robot. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 779–786. IEEE.

[202] Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. (2017). icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

[203] Renier, L. A., Anurova, I., De Volder, A. G., Carlson, S., VanMeter, J., and Rauschecker, J. P. (2009). Multisensory integration of sounds and vibrotactile stimuli in processing streams for "what" and "where". *Journal of Neuroscience*, 29(35):10950–10960.

[204] Repovš, G. and Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience*, 139(1):5–21.

[205] Rey, D. and Neuhäuser, M. (2011). *Wilcoxon-Signed-Rank Test*, pages 1658–1659. Springer Berlin Heidelberg, Berlin, Heidelberg.

[206] Richardson, D. C., Dale, R., and Kirkham, N. Z. (2007). The art of conversation is coordination. *Psychological science*, 18(5):407–413.

[207] Robins, A. (1995). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

[208] Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., and Wayne, G. (2019). Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32.

[209] Romano, J. M., Brindza, J. P., and Kuchenbecker, K. J. (2013). Ros open-source audio recognizer: Roar environmental sound detection tools for robot programming. *Autonomous robots*, 34:207–215.

[210] Romero-Garcés, A., Calderita, L. V., Martınez-Gómez, J., Bandera, J. P., Marfil, R., Manso, L. J., Bustos, P., and Bandera, A. (2015). The cognitive architecture of a robotic salesman. In *Proceedings of the Conferencia de la Asociación Española para la Inteligencia Artificial CAEPIA*, volume 15, pages 16–24.

[211] Roskies, A. L. (1999). The binding problem. *Neuron*, 24(1):7–9.

[212] Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., and Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In *2008 IEEE International Conference on Robotics and Automation*, pages 962–967. IEEE.

[213] Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. (2016). Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

[214] Röhl, M. and Uppenkamp, S. (2012). Neural coding of sound intensity and loudness in the human auditory system. *Journal of the Association for Research in Otolaryngology*, 13(3):369–379.

[215] Saaty, T. L. (2007). Time dependent decision-making; dynamic priorities in the ahp/anp: Generalizing from points to functions and from real to complex variables. *Mathematical and Computer Modelling*, 46(7-8):860–891.

[216] Samsonovich, A. V. (2012). On a roadmap for the bica challenge. *Biologically Inspired Cognitive Architectures*, 1:100–107.

[217] Sandini, G., Mohan, V., Sciutti, A., and Morasso, P. (2018). Social cognition for human-robot symbiosis—challenges and building blocks. *Frontiers in neurorobotics*, 12:34.

[218] Santucci, V. G., Oudeyer, P.-Y., Barto, A., and Baldassarre, G. (2020). Intrinsically motivated open-ended learning in autonomous robots.

[219] Schilbach, L., Wilms, M., Eickhoff, S. B., Romanzetti, S., Tepest, R., Bente, G., Shah, N. J., Fink, G. R., and Vogeley, K. (2010). Minds made for sharing: initiating joint attention recruits reward-related neurocircuitry. *Journal of cognitive neuroscience*, 22(12):2702–2715.

[220] Schweizer, K. and Moosbrugger, H. (2004). Attention and working memory as predictors of intelligence. *Intelligence*, 32(4):329–347.

[221] Sciutti, A., Barros, P., Castellano, G., and Nagai, Y. (2022). Affective shared perception. *Frontiers in Integrative Neuroscience*, 16.

[222] Sciutti, A., Patane, L., Nori, F., and Sandini, G. (2014). Understanding object weight from human and humanoid lifting actions. *IEEE Transactions on Autonomous Mental Development*, 6(2):80–92.

[223] Seth, A. K., McKinstry, J. L., Edelman, G. M., and Krichmar, J. L. (2004). Visual binding through reentrant connectivity and dynamic synchronization in a brain-based device. *Cerebral Cortex*, 14(11):1185–1199.

[224] Shams, L., Kamitani, Y., and Shimojo, S. (2000). What you see is what you hear. *Nature*, 408(6814):788–788.

[225] Shams, L., Kamitani, Y., and Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive brain research*, 14(1):147–152.

[226] Shimoda, S., Jamone, L., Ognibene, D., Nagai, T., Sciutti, A., Costa-Garcia, A., Oseki, Y., and Taniguchi, T. (2022). What is the role of the next generation of cognitive robotics? *Advanced Robotics*, 36(1-2):3–16.

[227] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556.*

[228] Siposova, B. and Carpenter, M. (2019). A new look at joint attention and common knowledge. *Cognition*, 189:260–274.

[229] Steel, A., Billings, M. M., Silson, E. H., and Robertson, C. E. (2021). A network linking scene perception and spatial memory systems in posterior cerebral cortex. *Nature communications*, 12(1):2632.

[230] Stein, L. A. (1997). Postmodular systems: Architectural principles for cognitive robotics. *Cybernetics & Systems*, 28(6):471–487.

[231] Sun, R. (2006). The clarion cognitive architecture: Extending cognitive modeling to social simulation. *Cognition and multi-agent interaction*, pages 79–99.

[232] Taatgen, N. and Anderson, J. R. (2010). The past, present, and future of cognitive architectures. *Topics in Cognitive Science*, 2(4):693–704.

[233] Teufel, C., Alexis, D. M., Todd, H., Lawrance-Owen, A. J., Clayton, N. S., and Davis, G. (2009). Social cognition modulates the sensory coding of observed gaze direction. *Current Biology*, 19(15):1274–1277.

[234] Thellman, S. and Ziemke, T. (2020). Do you see what i see? tracking the perceptual beliefs of robots. *Iscience*, 23(10):101625.

[235] Thórisson, K. R. (1999). Mind model for multimodal communicative creatures and humanoids. *Applied Artificial Intelligence*, 13(4-5):449–486.

[236] Toffa, O. K. and Mignotte, M. (2020). Environmental sound classification using local binary pattern and audio features collaboration. *IEEE Transactions on Multimedia*, 23:3978–3985.

[237] Tomasello, M. (2009). *The cultural origins of human cognition*. Harvard university press.

[238] Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–691.

[239] Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

[240] Triesch, J., Teuscher, C., Deák, G. O., and Carlson, E. (2006). Gaze following: Why (not) learn it? *Developmental science*, 9(2):125–147.

[241] Ude, A., Wyart, V., Lin, L.-H., and Cheng, G. (2005). Distributed visual attention on a humanoid robot. In *5th IEEE-RAS International Conference on Humanoid Robots, 2005.*, pages 381–386. IEEE.

[242] Ungerleider, L. G. and Haxby, J. V. (1994). 'what'and 'where'in the human brain. *Current opinion in neurobiology*, 4(2):157–165.

[243] van de Ven, G. M., Siegelmann, H. T., and Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):1–14.

[244] Vanrullen, R. and Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *Journal of cognitive neuroscience*, 13(4):454–461.

[245] Verwimp, E., De Lange, M., and Tuytelaars, T. (2021). Rehearsal revealed: The limits and merits of revisiting samples in continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9385–9394.

[246] Viciana-Abad, R., Marfil, R., Perez-Lorenzo, J. M., Bandera, J. P., Romero-Garces, A., and Reche-Lopez, P. (2014). Audio-visual perception system for a humanoid robotic head. *Sensors*, 14(6):9522–9545.

[247] Vignolo, A., Noceti, N., Rea, F., Sciutti, A., Odone, F., and Sandini, G. (2017). Detecting biological motion for human–robot interaction: A link between perception and action. *Frontiers in Robotics and AI*, page 14.

[248] Wagner, A. R. (1979). Habituation and memory. *Mechanisms of learning and motivation: A memorial volume for Jerzy Konorski*, pages 53–82.

[249] Wallach, H. (1939). On sound localization. *Journal of the Acoustical Society of America*, 10:270–274.

[250] Wallach, H. (1940). The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339 – 368.

[251] Warren, Z. E., Zheng, Z., Swanson, A. R., Bekele, E., Zhang, L., Crittendon, J. A., Weitlauf, A. F., and Sarkar, N. (2015). Can robotic interaction improve joint attention skills? *Journal of autism and developmental disorders*, 45(11):3726–3734.

[252] Weng, J. (2002). A theory for mentally developing robots. In *Proceedings 2nd International Conference on Development and Learning. ICDL 2002*, pages 131–140. IEEE.

[253] Wilson, J. and Lin, M. C. (2020). Avot: Audio-visual object tracking of multiple objects for robotics. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10045–10051. IEEE.

[254] Yamakawa, N., Takahashi, T., Kitahara, T., Ogata, T., and Okuno, H. G. (2011). Environmental sound recognition for robot audition using matching-pursuit. In *Modern Approaches in Applied Intelligence: 24th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2011, Syracuse, NY, USA, June 28–July 1, 2011, Proceedings, Part II 24*, pages 1–10. Springer.

[255] Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., et al. (2018). The grand challenges of science robotics. *Science robotics*, 3(14):eaar7650.

[256] Yoon, J., Jeong, W., Lee, G., Yang, E., and Hwang, S. J. (2021). Federated continual learning with weighted inter-client transfer. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12073–12086. PMLR.

[257] Yu, C. and Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLOS ONE*, 8(11):e79659.

[258] Zaadnoordijk, L., Besold, T. R., and Cusack, R. (2022). Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence*, 4(6):510–520.

[259] Zhong, X., Sun, L., and Yost, W. (2016). Active binaural localization of multiple sound sources. *Robotics and Autonomous Systems*, 85:83–92.

[260] Zmigrod, S. and Hommel, B. (2013). Feature integration across multimodal perception and action: a review. *Multisensory research*, 26(1-2):143–157.