Biomimetic Computation and Embodied Embedded Cognition for Spatial Audition in Humanoids

Dissertation

with the aim of achieving the degree of
Doctor rerum naturalium (Dr. rer. nat.) at the
Faculty of Mathematics, Informatics and Natural Sciences,
Department of Informatics,
University of Hamburg.

Submitted by
Jorge Dávila Chacón
2019 in Hamburg, Germany.

The following evaluators recommend the admission of the dissertation:

Dr. Jindong Liu
Department of Computing
Imperial College London, UK

Prof. Dr. Jianwei Zhang
Department of Informatics
University of Hamburg, Germany

Prof. Dr. Timo Gerkmann
Department of Informatics
University of Hamburg, Germany

Prof. Dr. Frank Steinicke (chair)
Department of Informatics
University of Hamburg, Germany

Prof. Dr. Stefan Wermter (advisor)
Department of Informatics
University of Hamburg, Germany

Submitted on:
9^{th} of March, 2019
Day of oral defence:
30^{th} of April, 2019

To my family:
Soco, Nacho, Tita and Pia.

Abstract

Inspired by the behaviour of humans talking in noisy environments, we propose an embodied embedded cognition approach to improve automatic speech recognition (ASR) for robots under challenging conditions, such as high levels of ego-noise, using binaural sound source localisation (SSL). We find that the humanoid embodiment allows the generation of additional spatial cues that cover the entire audible range, without additional computational costs. Furthermore, by simplifying existing biomimetic models for the extraction of spatial cues in sound, we are able to understand the principles that are important to perform robustly in noisy environments. We test our approach by measuring the impact of SSL with a humanoid robot head on the performance of an ASR system. More specifically, the robot orients towards the angle where the signal-to-noise ratio (SNR) of speech is maximised for one microphone and uses this signal as input to the ASR system. In our first experiment, we make use of one humanoid platform (Nao) to produce the spatial cues necessary for SSL. The embodiment of the robot produces cues that are robust to interfering noise as they span a broad range of sound frequencies. Then, we use spiking neural networks (SNN) to extract such spatial cues from the sound. The SNN are biomimetic models of regions in the mammalian midbrain that are relevant for SSL. Next, a Bayesian model integrates the spatial cues encoded by the biomimetic models and a feedforward neural network is used to handle high levels of ego-noise and reverberation in the signal. Once the robot determines the direction of the incoming sound, it turns in the direction of the sound source, and the sound signal is fed into an ASR system. For ASR, we use DOCKS, a system developed by the Knowledge Technology Group of the University of Hamburg, and compare its performance with and without support from the SSL system. In order to measure the quality of the spatial cues created by different robot embodiments, we test our SSL and

ASR systems on two humanoid platforms with different structural and material properties (iCub and Soundman). With our approach, we halve the sentence error rate in comparison to the standard approach of downmixing the input of both channels. We find that ASR performs more than two times better when the angle between the humanoid head and the sound source allows sound waves to be reflected most intensely from the pinna to the ear microphone, rather than when sound waves arrive perpendicularly to the membrane. In conclusion, our work allows understanding in greater detail the advantages of using a humanoid embodiment to produce spatial cues and of using biomimetic models to represent such cues. Equally important, we also understand better the importance of robots that use behaviour as a programmatic approach that converges in a sequence of steps to the optimal configuration for performing ASR in noisy conditions.

Keywords: Automatic speech recognition, behavioural robotics, binaural sound source localisation, bioinspired neural architectures.

Zusammenfassung

Menschen sind besonders gut darin, sich in geräuschvollen Umgebungen zu unterhalten. Davon inspiriert, schlagen wir einen kognitiven, in körperliche Wahrnehmung eingebetteten Ansatz zur Verbesserung von automatischen Spracherkennungssystemen (ASR) vor. Dieser Ansatz ermöglicht die ASR auf Robotern unter besonders schwierigen Bedingungen, beispielsweise unter Egogeräuschen, unter Zuhilfenahme von binauraler Geräuschquellenlokalisierung (SSL). Wir überprüfen unseren Ansatz, indem wir die Auswirkung von SSL in der Performanz eines ASR-Systems mit einem humanoiden Roboterkopf bemessen. Insbesondere wird dem Roboter ermöglicht, sich in die Richtung des Winkels zu orientieren, in welchem das Signal-Rausch-Verhältnis (SNR) von natürlicher Sprache für ein Mikrophone am Besten ist und dann dieses Signal als Eingabe für das ASR-System zu benutzen. Zuerst machen wir uns dabei eine humanoide Plattform zu Nutze um

räumliche Hinweise zu erzeugen, die notwendig für die SSL sind. Als nächstes benutzen wir gepulste neuronale Netzwerke (SNN), um diese räumlichen Hinweise aus dem Sound zu extrahieren. Die SSN sind bio-mimetische Modelle für Regionen im Mittelhirn von Säugetieren, welche als besonders relevant für die SSL angesehen werden. Schließlich integrieren wir mit einem Bayesischen Modell die räumlichen Hinweise, welche von den bio-mimetischen Modellen enkodiert werden, und benutzen ein neuronales Feedforward-Netzwerk um den hohen Grad an Egogeräuschen und Widerhall des Sounds zu bewältigen. Nachdem der Roboter die Richtung des eingehenden Sounds bestimmt hat, dreht sich dieser in die Richtung der Soundquelle und speist das Sound-Signal in das ASR-System ein. Für die ASR benutzen wir ein System, welches eigens in unsere Gruppe entwickelt wurde und vergleichen damit die Performanz, sowohl mit als auch ohne die Unterstützung unseres SSL Ansatzes. Um die Qualität von räumlichen Hinweisen zu bemessen, die sich aus eingebetteten Körperwahrnehmungen unterschiedlicher Roboter ergeben, untersuchen wir unseren SSL- und ASR-Systeme auf zwei humanoiden Roboterplattformen mit unterschiedlichen Struktur- und Materialeingenschaften. Mit unserem Ansatz sind wir in der Lage, die Fehlerrate auf Sätzen zu halbieren, verglichen mit dem Standardansatz, bei dem die Eingabe aus zwei Kanälen heruntergemischt wird. Wir finden, dass das ASR-System mehr als zweifach besser funktioniert, wenn der Winkel zwischen dem humanoiden Kopf und der Soundquelle es ermöglicht, dass die Soundwellen am intensivsten von der Ohrmuschel zum Mikrophon des Ohres reflektiert werden, anstatt wenn die Soundwellen senkrecht auf die Membran auftreffen. Zusammengefasst, ermöglicht unsere Arbeit sowohl ein tieferes Verständnis über die Möglichkeiten, wie wir humanoide eingebettete Körperwahrnehmung nutzen können, um räumliche Hinweise zu erzeugen, als auch, wie wir bio-mimetische Modelle zur deren Repräsentation einsetzen können. Gleichermaßen wichtig ist auch unser verbessertes Verständnis über die Wichtigkeit für Roboter, ein Verhalten als programmatische Annäherung zu nutzen, welches in einer Abfolge von Schritten zur optimalen Konfiguration konvergiert, um ASR unter geräuschvollen Bedingungen zu leisten.

Keywords: Automatische Spracherkennung, Verhaltensrobotik, binaurale Schallquellenlokalisierung, bioinspirierte neurale Strukturen.

Contents

Glossary

S	Capital letters indicate sets.
$\in$	Set membership.
$\lvert \ldots \rvert$	Cardinality of a set.
$\mathbf{M}$	Boldface capital letters indicate 2D arrays.
$\odot$	Element-wise array multiplication.
$\forall$	Universal quantification.
$\wedge$	Logical conjunction.
$\vee$	Logical disjunction.
$\neg$	Logical negation.
$[,]$	Closed interval.
$\sim$	Same order of magnitude.
$\gg$	Of greater order than.
$\lvert$	Conditional event.
$O()$	Computational complexity.

Chapter 1

Introduction

Sound conveys information that is crucial for our interaction with the environment. This information is particularly useful when the environment obstructs visual information, e.g., when the light is scarce or in environments cluttered by the presence of dense vegetation, fog, etc. Sound not only conveys information about the occurrence of a given event in time and space (Griffiths & Warren, 2004) but also about its context (Hengel & Andringa, 2007), the relation between different events and the physical properties of materials (Sinapov *et al.*, 2011). Therefore, audition allows us to create a more accurate and dynamic representation of the world, which is essential for the emergence of intelligence (McCarthy, 1960; Minsky, 1961; Newell *et al.*, 1972; McCarthy & Hayes, 1981; Samsonovich, 2012). Audition is a broad field of study, and in the present work we focus on the extraction of spatial information contained in sound. This subfield of auditory perception is known as sound source localisation (SSL). SSL is an essential ability for animals to survive, as the continuous spatial localisation of a sound source can inform the listener about the dynamics of the world, e.g., the direction and speed of multiple sound sources. SSL can be useful in a wide range of behaviours in nature, including competition strategies like the detection of predators and the accurate targeting of prey (Kim, 2006). Localising sounds in space can also be crucial for mating, communication and in general for survival.

More specifically, we are interested in the auditory system of humans (Wright & Zhang, 2006). People routinely display behaviours that are important for interacting with dynamic environments. This range of conducts is made possible

by our internal representation of the world acquired through our senses and integrated by our brains (Bowers, 2009; Kourtzi & Connor, 2011; West *et al.*, 2018). This integrative process is called perception, and it is a complex cognitive function that allows humans to create such representations and find meaning in them.

Even though the information we receive is subject to noise from several sources, the integration of different sensory modalities can provide the necessary redundancy to perceive the environment with consistency (Stein & Meredith, 1993a; Dosher & Lu, 1998; Ernst & Bülthoff, 2004; Hartmann *et al.*, 2005). In the case of auditory perception, our brain extracts various types of information contained in sound. The first layers in our auditory pathway extract low-level features of sound. These initial stages of auditory processing allow us to segregate individual sound components from noisy backgrounds, localise them in space and detect their motion patterns (Lopez-Poveda *et al.*, 2010; Ruggles *et al.*, 2011; Moore, 2012; Grothe, 2000; Grothe *et al.*, 2010). In later stages, our brain extracts high-level auditory features to perform tasks such as understanding natural language (Schnupp *et al.*, 2011; Golumbic *et al.*, 2013).

For all the previous reasons, audition is also crucial for autonomous robotic systems (van der Zant & Iocchi, 2011; Stramadinoli *et al.*, 2011; Andersson *et al.*, 2004). Notably, the ability to pinpoint sound sources is essential for the safe interaction of robots with the environment and for improving communication with humans (Roman *et al.*, 2003). Its azimuth, elevation and depth specify the location of a sound source in space. However, it is only possible for a listener to estimate the distance to a sound source when the nature of the sound is familiar to the listener (Nakashima & Mukai, 2005; Schenkman & Nilsson, 2011). For example, we can estimate how far is our dog when it barks, because it always does it with the same intensity. In this project, we focus on sound source localisation on the frontal 180° along the azimuth plane, as our focus is on Human-Robot Interaction, i.e., on tracking the voice of the speaker that the robot is facing. Furthermore, we also investigate the use of spatial cues in sound to improve automatic speech recognition (ASR), as the spatial localisation of a speaker on the azimuth can increase the signal-to-noise ratio in *Cocktail Party* scenarios and support high-level cognitive tasks (Roman *et al.*, 2003; Delcroix *et al.*, 2011;

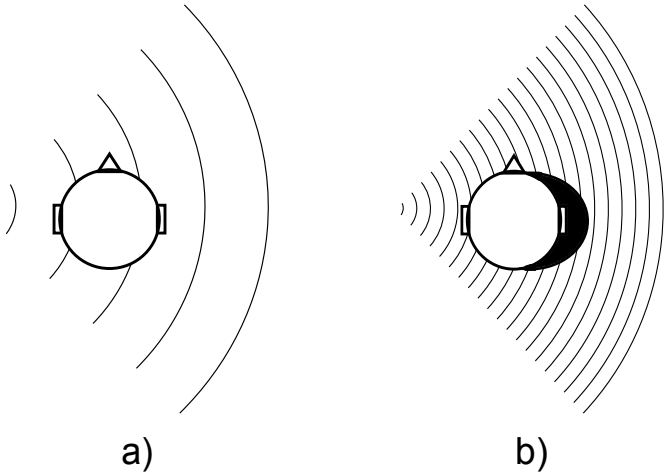


Figure 1.1: **a)** Interaction of a head structure and low-frequency components in sound. **b)** Interaction of a head structure and high-frequency components in sound. Notice that the head produces a considerable shadowing effect only with high frequencies (Blauert, 1997, Ch. 2.2.2).

Hurmalainen *et al.*, 2011; Marti *et al.*, 2012; Hill *et al.*, 2012; Spille *et al.*, 2013; Jiang & Liu, 2014).

As with any other perceptual capability, a meta-objective of artificial SSL systems is their portability between different robotic platforms (Yamamoto *et al.*, 2004). This meta-objective partly explains the broad range of approaches that scientific literature has documented, including complex microphone arrays fitted to specific rooms and robotic platforms. An alternative paradigm to multiple microphone arrays is binaural SSL, as humans are a clear example that it is possible to achieve accurate sound source localisation using only two sound sensors or ears. Humans rely on the effect produced by the pinnae, head and torso on the sound frequency components (FC), and on the capacity to move our head for performing SSL (Middlebrooks & Green, 1991). Similarly, with only one pair of microphones separated by a head-like structure, an SSL system can estimate interaural time differences (ITD) and interaural level differences (ILD). Both spatial cues are fundamental, as ITDs convey more accurate information in low FCs and ILDs in high FCs. All these neurophysiological findings of sound source localisation in mammals inspired the scientific community to design novel systems for SSL during the last decade.

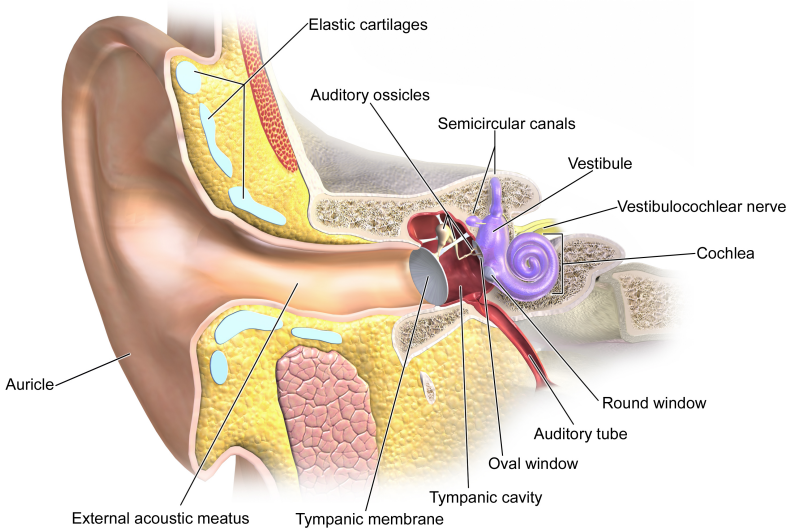


Figure 1.2: Anatomy of the entire human ear. Image from Wikimedia Commons. Freely distributed under the Creative Commons *Attribution-Share Alike 3.0 Unported* license.

1.1 Embodiment and Neural Correlates

In this section, we present an overview of the biological principles found by neuroanatomical studies of the mammalian auditory pathway (King & Palmer, 1983; Masterton & Imig, 1984; Jenkins & Merzenich, 1984; Kayser *et al.*, 2005; Goodman & Brette, 2010; Brette, 2012). More specifically, we describe the interaction between the body of the human listener and the approaching sound waves, the transduction of mechanical vibrations in the inner ear to neural spikes and the spatial encoding of information contained in sound that takes place at subsequent layers in our brain (Panchev & Wermter, 2006).

1.1.1 Torso and Pinnae

Sound waves are affected when they interact with our bodies. This interaction modifies the frequency spectrum of sound reaching our ear canal in different

ways, depending on the spatial location of the sound source around our body. Low FCs, with a wavelength at least twice as long as the interaural distance, can produce ITDs that indicate the angle of incoming sound unambiguously (Schnupp *et al.*, 2011; Lund *et al.*, 1998). However, the ITD for high frequencies in sounds starts becoming ambiguous once the wavelength of high-frequency components is less than twice the interaural distance. For example, in human adults, ITDs become ambiguous at frequencies above 1600 Hz (Middlebrooks & Green, 1991). The torso and pinnae reflect with different intensities high FCs, and the head does not diffract them around the head, reducing the sound pressure level at the contralateral ear. Such influence on the sound waves has a "shadowing" effect that generates specific ILDs for different angles along the azimuth. Figure 1.1 shows the interaction between a head-like structure and different frequency components in sound. ITDs and ILDs are complementary cues, as they contain information from both extremes of the audible frequencies range. As ILDs and ITDs allow the localisation of a sound source in space, their integration is known as the *Duplex Theory* of sound source localisation (Middlebrooks & Green, 1991).

Figure 1.2 shows the anatomy of the human ear. The geometry and material of the pinna affect the intensity of individual frequencies in the sound spectra due to reflection and absorption (Hofman *et al.*, 1998; Pujol *et al.*, 2019). This effect allows the front-back disambiguation of sound sources. After the sound reaches the eardrum, the middle ear ossicles transfer the air pressure waves into the cochlear fluid. Figure 1.3 shows the anatomy of the middle ear. There, the surface ratio between the eardrum and the oval window is around 20:1. Together with the mechanical amplification produced by the ossicles, the total pressure increase can reach up to 26 dB, varying with different frequencies and individuals. Afterwards, the middle ear behaves as an impedance adapter; it transfers efficiently mechanical waves from gas (air) to liquid (cochlear fluid). Without it, our ears would reflect in the environment approximately 98% of the sound waves (Pujol *et al.*, 2019). Together, the influence of the pinna and ossicles on the sound spectra provides essential monaural clues that allow us to determine the location of sound sources on the elevation plane.

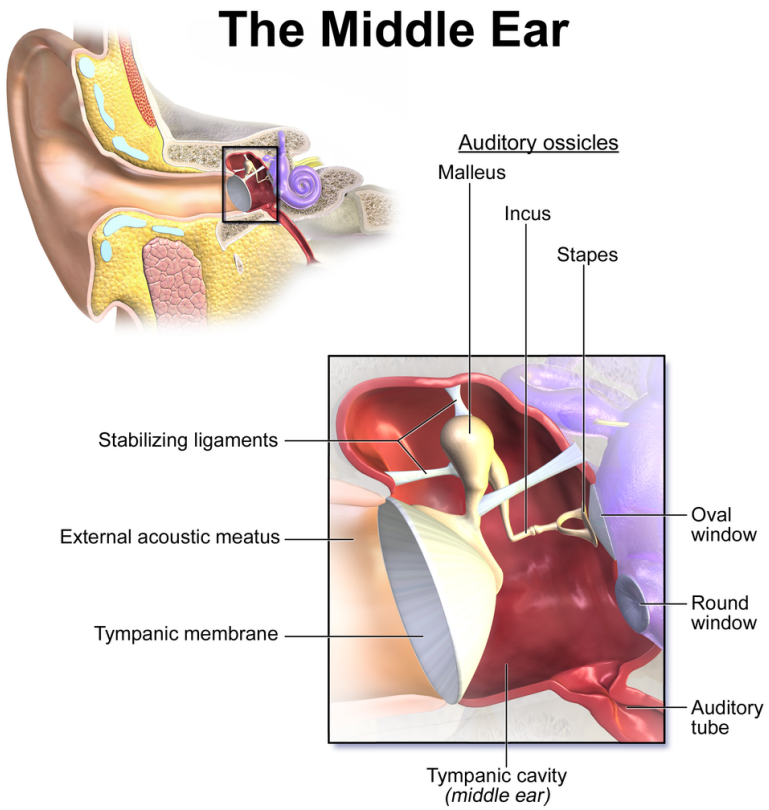


Figure 1.3: Anatomy of the human middle ear. Image from Wikimedia Commons. Freely distributed under the Creative Commons *Attribution-Share Alike 3.0 Unported* license.

1.1.2 Inner Ear

Figure 1.4 shows the anatomy of the inner ear. Once sound waves reach our inner ear, they produce vibrations inside the cochlea. The organ of Corti then encodes the information contained in these oscillation patterns by transducing mechanical vibrations on the basilar membrane (BM) into neural spikes (Richter *et al.*, 1998). Inside the cochlea, the BM functions like a mechanical filter that decomposes the sound wave in its fundamental frequencies. Such filtering is a clear example of the advantages of Embodied Embedded Cognition (Krichmar, 2012; Pfeifer *et al.*, 2007; Pulvermüller, 2013), as the passive mechanism of the BM performs this computation efficiently without the need for metabolism. Also inside the cochlea, the hair-cells (HC) transduce the mechanical vibrations along

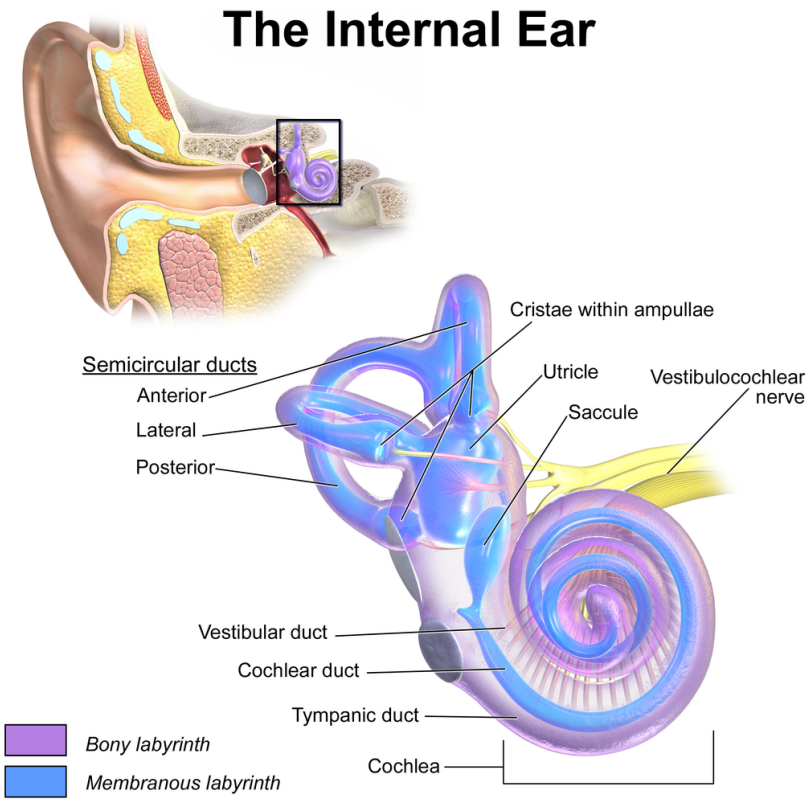


Figure 1.4: Anatomy of the human inner ear. Image from Wikimedia Commons. Freely distributed under the Creative Commons *Attribution-Share Alike 3.0 Unported* license.

the BM into neural spikes. These spikes are phase-locked to the section of the BM most sensitive to a particular frequency. The neural topology of the auditory pathway shows the same spatial distribution of FCs from the BM up to the auditory cortex Schnupp *et al.* (2011). Figures 1.6 and 1.7 show in detail the anatomy of the Cochlea and the Organ of Corti.

An HC has the highest probability of producing a spike when the local wave amplitude in the BM is maximal. As HCs are attached only to one side of the BM, they behave like a half-wave rectifier. Figure 1.5 shows waves representing vibrations in the left (L) and right (R) basilar membranes at a section resonant to a given sound frequency component f. The markers above the maximum amplitudes of the waves represent the point in time with the maximum probability

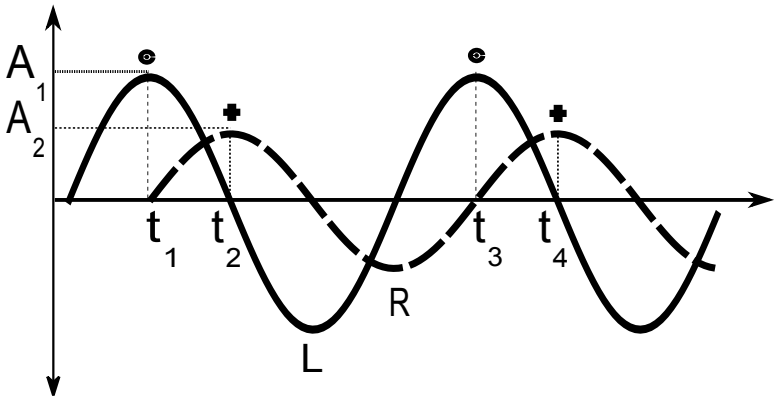


Figure 1.5: The waves represent vibrations on the left (L) and right (R) basilar membranes at sections that resonate with a given sound frequency component f. The markers above the maximum amplitudes of the waves represent the point in time with the maximum probability of a neural spike to be produced by the HCs in the organ of Corti.

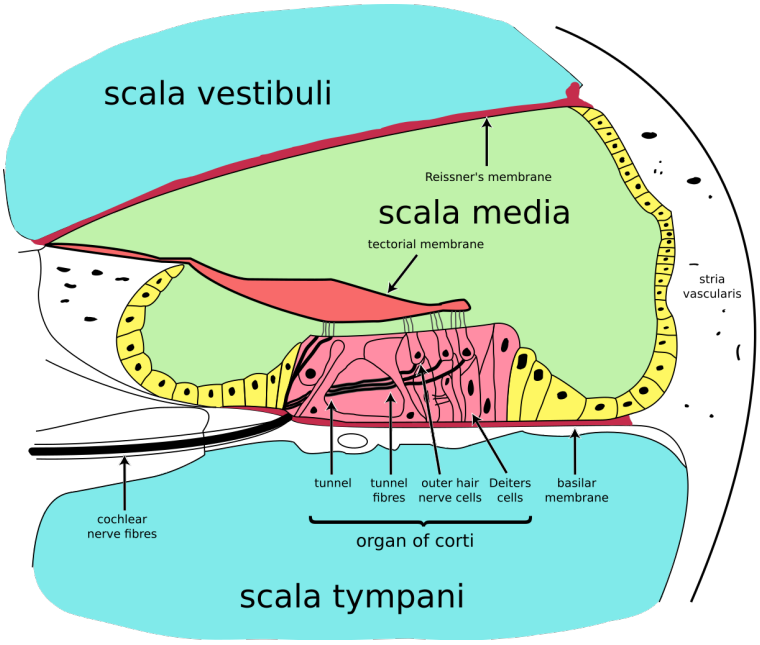


Figure 1.6: Cross section of the human Cochlea. Image from Wikimedia Commons. Freely distributed under the Creative Commons *Attribution-Share Alike 3.0 Unported* license.

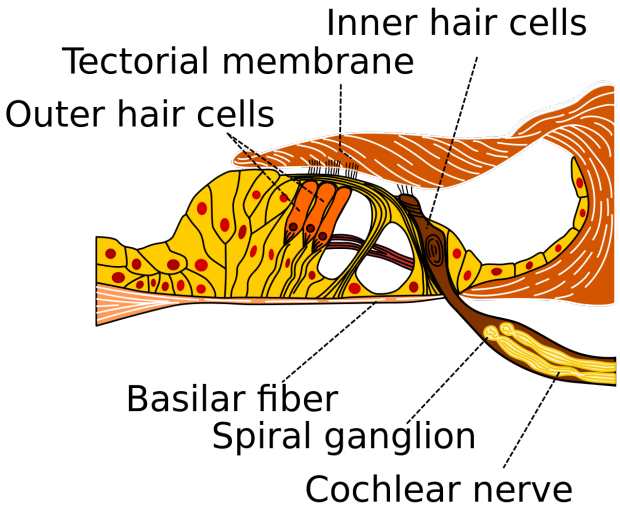


Figure 1.7: Anatomy of the Organ of Corti. Image from Wikimedia Commons. Freely distributed under the Creative Commons *Attribution-Share Alike 3.0 Unported* license.

of a neural spike to be produced by the HCs in the organ of Corti. Once stimulated, HCs release neurotransmitters to their corresponding fibres in the auditory nerve (AN). Each fibre of the AN has bifurcations to all the subdivisions of the cochlear nucleus (CN), the first relay station in the auditory pathway (Schnupp *et al.*, 2011). From the CN, different cell types convey temporal and spectral information to the medial superior olive (MSO) (Grothe, 2000; Oliver *et al.*, 2003; Roberts & Golding, 2012) and the lateral superior olive (LSO) respectively (Guinan *et al.*, 1972a,b; Park *et al.*, 2004). We are particularly interested in the MSO and LSO regions, as they extract ITDs and ILDs respectively.

1.1.3 Superior Olives and Inferior Colliculus

The MSO performs the task of a coincidence detector, where different neurones represent spatially different ITDs (Smith *et al.*, 1993; Biologie, 2007). Neurones in the MSO encode ITDs more effectively from the low-frequency components of sounds. Different delay mechanisms accomplished this representation, such as the different thickness of the axon myelin-sheaths, or different axon lengths from the excitatory neurones in the ipsilateral and contralateral cochlear nucleus (Joris *et al.*, 1998). Figure 1.8 presents the principle behind these mechanisms. In the

case of level differences, different neurones in the LSO represent spatially different ILDs (Glendenning & Masterton, 1983; Thompson & Dau, 2008; Brette, 2012). Due to the shadowing effect of the head, the LSO encodes ILDs more effectively from the high-frequency components of sound (Irvine *et al.*, 2001). The mechanism underlying the extraction of ILDs is not clearly understood in comparison to the mechanism of ITDs. Nevertheless, we know that LSO neurones receive excitatory input from the ipsilateral ear and inhibitory input from the contralateral ear. From this input, different neurones in the LSO display a typical spiking rate for sound sources located at specific angles along the azimuthal plane(Schnupp *et al.*, 2011). Precise inhibition is essential for microsecond interaural time difference (Brand *et al.*, 2002; Grothe, 2003; Vasilkov & Tikidji-Hamburyan, 2012).

In the following station in the auditory pathway, the inferior colliculus (IC) integrates the output of the MSO and LSO layers (Chase & Young, 2008; Escabi & Schreiner, 2002) and directs its output to cortical areas (Salminen *et al.*, 2010; Atencio *et al.*, 2012). Even though the IC receives forward connections from the peripheral areas and recurrent connections from the higher-level areas (thalamic and cortical), one of its main tasks is the integration of ITDs and ILDs into a coherent spatial representation of sound sources Recanzone & Sutter (2008); Andersson *et al.* (2004). We can think of the combination of both spatial cues as a multimodal integration process Stein (1967); Stein & Meredith (1993b), where ITDs and ILDs are the modalities to be integrated in order to sharpen the neural representation of sound sources in the environment. Finally, the scientific literature shows that thalamocortical areas can be relevant for SSL (Recanzone & Sutter, 2008; Huo & Murray, 2009). However, the exact dynamics of such influence remain unclear, and therefore we do not consider it in this work.

1.2 Research Objectives

From a *global perspective*, we consider the objectives of the research framework of the International Graduate Research Group on Cross-Modal Interaction in Natural and Artificial Cognitive Systems (CINACS)[1] to provide a framework for

[1] `https://cinacs.informatik.uni-hamburg.de/about-cinacs`

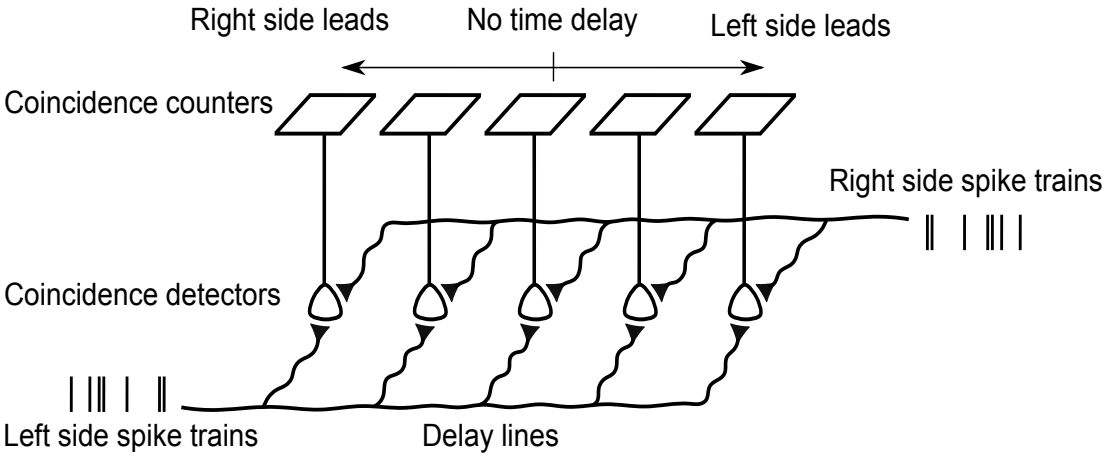


Figure 1.8: Diagram of the MSO modelled as a Jeffress coincidence detector for representing ITDs (Jeffress, 1948). This model compares spikes produced by the same frequency components f when the time difference δt between spikes is smaller than half a period. This is, when $2f \cdot \delta t < 1$.

the present work. Our *guiding hypothesis*, is that embodiment and cross-modal integration provide the necessary basis to develop the next generation of artificial cognitive systems (Krichmar, 2012; Stork, 2012; Hiatt *et al.*, 2012; Winston, 2012; Kelso *et al.*, 2013). The importance of these two principles resides in the extraction of information from the environment through embodiment, and in the integration diverse sources of information to facilitate a more robust representation of the world (Koch, 1993; Wilson, 2002; Metta *et al.*, 2008; Pulvermüller, 2013). With the integration of biological and engineering approaches, we intend to generate complementary knowledge in both fields in a continuous cycle (Wermter *et al.*, 2005), rather than only focusing in the direction of reverse-engineering (Schierwagen, 2012). CINACS promoted the continuous interaction between research groups in diverse disciplines including us, the Knowledge Technology Group. During such exchanges, specific research questions provided a framework for our discussions around cross-modal interactions and defined our approach to understand spatial cognition, e.g., in some cases, what seems to be purely visual phenomena can be better understood with the involvement of auditory phenomena, and vice versa (Shinn-Cunningham, 2008). It is important to clarify that, although our system only works with one sensory input, we treat

the multiple spatial cues as information encoded in sound (Shannon, 1948) as separate modalities that can be integrated to provide a richer and more accurate representation of the world. Hence, we try to answer the following questions:

- Which architectures are suitable for certain types of cross-modal tasks?
- How to transform between modalities?
- What are the mechanisms of cross-modal perceptual phenomena?
- What are the general principles for resolving cross-modal conflicts?
- How are multimodal percepts generated and represented?
- How can cross-modal integration be realised in technical systems?

From a *concrete perspective*, the objective of this work is to gain insights about the bottom-up and top-down influence of embodiment for spatial audition in natural and artificial systems. As documented in this thesis, we have designed an architecture to improve robot speech recognition, based on the principles of biomimetic computation and embodied embedded cognition. In this context, we have adapted some of the CINACS objectives[1] to determine the guidelines that directed our experimental work:

1. To improve our understanding of acoustic localisation through cross-modal integration.
2. To understand acoustic localisation from an integrated view of spatial audition at multiple scales.
3. To introduce biological principles into artificial intelligent systems for acoustic localisation.

Our *first objective*, is to increase our understanding of the influence of humanoid embodiment on bottom-up cognitive tasks for sound perception (Koch, 1993; Hofman *et al.*, 1998; Horimoto *et al.*, 2012), such as static and dynamic SSL. The first step is the selection of the robotic platforms for our experimental

setup. If the best interface for a human is another human (Wilson, 2002), we should exploit the computational advantages that embodiment brings "for free". In the present work we use three robotic platforms: Nao (Gouaillier *et al.*, 2009), iCub (Beira *et al.*, 2006) and Soundman (Salb & Duhr, 2009). Both, Nao and iCub, are humanoid robots designed for research in academia, and Soundman is a platform designed for binaural recordings that maximise the generation of sound spatial cues. As the design of the iCub robot is intended for research in Cognitive Developmental Robotics (Metta *et al.*, 2008), it approaches the physiognomy of humans and allows to measure more precisely the influence of a humanoid embodiment on our models of the auditory system. In the present work, we are not interested in the design of a generic SSL or ASR system with higher accuracy than existing systems.

Our *second objective*, is to increase our understanding about the influence of embodiment on top-down cognitive tasks (Koch, 1993; Zhao *et al.*, 2018) like ASR, when using biomimetic models of bottom-up cognition like SSL (Singheiser *et al.*, 2012). There is ample literature about robotic ASR, including systems that perform SSL with large microphone arrays to improve ASR. However, we are one of the first and few groups working on SSL and ASR inside the framework of embodied embedded cognition (Finger & Liu, 2011). This circumstance reduces the amount of scientific literature available for a comparison of different methodologies (Wilson, 2002; Nguyen *et al.*, 2018), but at the same time, it highlights the need to expand our understanding in this direction. Once the behaviour of the robot corresponds to the behaviour of animals (Noë & Regan, 2000; Nodal *et al.*, 2010; Greene *et al.*, 2012), we can observe the activity of the neural models under new conditions and produce new hypothesis to guide further studies in biological systems, such as studies in human speech recognition (HSR).

Our *third objective*, is to close the loop by using the experimental results obtained with artificial systems to guide further research in natural systems (van Hateren, 1992; Barrès *et al.*, 2013; Famulare & Fairhall, 2010). As pointed out by Scharenborg (2007), further research is necessary to understand better the auditory cues used by human listeners, and that possibly are being overlooked in current artificial systems. Once these features (acoustic or from other sensory modalities) are recognised, researchers can readily integrate them into the

design of novel multimodal architectures (Benoit *et al.*, 2000; Schauer & Gross, 2003; Goertzel *et al.*, 2010). More specifically, Scharenborg asks how can such knowledge about child language acquisition be used to improve ASR systems and computational models of HSR? He proceeds then to conjecture that understanding how infants acquire language could lead to the design of new paradigms for ASR, well beyond the probabilistic pattern recognition techniques that modern systems commonly use. One example being when children acquire language. At this developmental stage the units for the segmentation of acoustic signals are not pre-specified, as is nowadays the case for ASR systems and computational models of HSR. In order to achieve such flexibility, it is necessary to develop novel architectures that make use of emergent units of recognition, instead of constraining the systems to use the linguistic units present in current ASR systems and computational models.

1.3 Novel Contribution to the Field

The objectives defined in Section 1.2 are tightly coupled; therefore our experiments have not addressed each of them separately, but conjunctly. Concerning objective 1, we have improved our understanding of the neural mechanisms used for the integration of sound spatial cues in mammalian brains (Glackin *et al.*, 2010; Fischer & Peña, 2011; Fontaine & Brette, 2011). More specifically, it has become clear that the topology of connections between layers in the auditory pathway can improve the signal-to-noise ratio of information transmitted to the higher layers (See Section 2.2). As we can interpret the topological constraints found in natural systems as hyperparameters in computational models, it is then possible to implement such constraints in biomimetic architectures. We can then proceed to measure their accuracy by replicating ethological experiments with robots, and measure their predictive power by observing their behaviour in previously unseen scenarios. For this particular purpose we have designed a virtual reality experimental setup designed for audio-visual integration (Bauer *et al.*, 2012). This setup allows us to measure the response of the system to controlled stimuli, at the neural and behavioural levels, with high precision. These accom-

plishments are in line with objectives 1 and 2. We provide a detailed description of the virtual reality setup in Section 5.1.1.

The biological principles that we have introduced into an artificially intelligent system (objective 3) range from the computation performed by the embodiment of the robot itself, to the biomimetic computational models used to filter and encode the signals sensed by the robot. Particularly after our last experiment (Chapter 5.3), we gained insights into the computation performed by the asymmetrical absorption of sound frequencies with the humanoid pinnae. Another important insight is the benefit of the efficient computation performed in the inner ear. There, the Organ of Corti performs the mechanical transduction of vibrations in the basilar membrane without requiring additional metabolism, i.e., without the need for consuming additional energy resources for quasi-instantaneous computation. The results of the experiments presented in this work have increased our understanding of the improvements achieved by the generation of spatial cues with a humanoid head, and the benefits of constraining the search space of hyperparameters by following anatomical guidelines found in biological systems (Chapter 3).

1.3.1 Publications Originating from this Thesis

The present work produced the following publications during its development:

(I) Journals:

(1) J. Bauer, J. Davila-Chacon, S. Wermter. Modelling the development of natural multi-sensory integration using neural self-organisation and probabilistic population codes. Connection Science, 2014.

(2) J. Davila-Chacon, J. Liu, S. Wermter. Enhanced Robot Speech Recognition Using Biomimetic Binaural Sound Source Localisation. IEEE Transactions on Neural Networks and Learning Systems, 2018.

(II) Conferences:

(3) J. Bauer, J. Davila-Chacon, E. Strahl, S. Wermter. Smoke and Mirrors—Virtual Realities for Sensor Fusion Experiments in Biomimetic

Robotics. IEEE International Conference on Multisensor Fusion and Information Integration (ICMF), Hamburg, Germany, 2012.

(4) J. Davila-Chacon, S. Heinrich, J. Liu, S. Wermter. Biomimetic Binaural Sound Source Localisation with Ego-Noise Cancellation. International Conference on Artificial Neural Networks (ICANN), Lausanne, Switzerland, 2012.

(5) J. Davila-Chacon, S. Magg, J. Liu, S. Wermter. Neural and Statistical Processing of Spatial Cues for Sound Source Localisation. International Joint Conference on Neural Networks (IJCNN), Dallas, USA, 2013.

(6) J. Davila-Chacon, J. Twiefel, J. Liu, S. Wermter. Improving Humanoid Robot Speech Recognition with Sound Source Localisation. International Conference on Artificial Neural Networks (ICANN), Hamburg, Germany, 2014.

(III) Abstracts:

(7) J. Davila-Chacon. Neural Sound Source Localisation for Speech Processing Based on the Inferior Colliculus. In Proceedings of the Joint Workshop of the German Research Training Groups in Computer Science, 2012, 2013 and 2014.

1.4 Thesis Organisation

Chapter 1 introduces the topics from animal neurophysiology that are relevant to the biomimetic computational model that we use for SSL and Chapter 2 provides an overview of the evolution of artificial SSL systems. It starts with an overview of the initial approaches using large microphone arrays, followed by the second generation robotic approaches and concluding with an overview of the more recent bioinspired architectures. In particular, section 2.2 explains how we adapted this knowledge to the context of robots producing ego-noise. Such adaptations include a simplified version of the spiking neural network and the Bayesian model that we use as a starting point to integrate multiple spatial cues.

Then the following chapters then introduce our experimental work. Chapter 3 details the importance of optimising the hyperparameters that determine the measurement of interaural level differences and explains how they are dependent on the geometry of the robotic head. Chapter 4 reflects one of the most significant contributions of the present work, as it explores the advantages of combining neural and statistical methods to achieve the required balance between life-long learning and computational costs. Chapter 5 integrates our work in SSL with the field of automatic speech recognition (ASR). As mentioned before, a pervasive challenge in the field of robotics is the addition of high levels of ego-noise produced by the cooling systems. Our objective in the two experiments that we present in the last chapter is to measure the improvement of ASR when we combine it SSL. Interestingly, ASR performs best when the angle between the humanoid head and the sound source allows sound waves to be reflected most intensely from the pinna to the ear microphone, rather than when sound waves arrive perpendicularly to the membrane. The first experiment in Section 5.1.2 explores the effect of the embodiment of two robotic platforms. The second experiment in Section 5.3 concludes our journey by studying the interaction between the robotic platform and the sound source, i.e., we analyse the effect on ASR of turning towards a human speaker in different locations inside and outside of the visual field of view. Finally, Chapter 6 summarises the results that we obtain in our empirical studies and elaborates on the answers that they provide to our research objectives.

Chapter 2

Development of Computational Methods

During the last decade, plenty of neurophysiological findings related to sound source localisation in mammals inspired the scientific community to design bioinspired systems for SSL. In order to contextualise the contribution of the present work, this chapter outlines the most representative methods used for robotic SSL in the past three decades. The objective is to understand the importance of SSL as a technology that can support complex devices, such as robots, but also to understand its importance as a window for observing some fundamental aspects of human cognition. A historical perspective also reveals the most significant challenges that SSL systems have faced and the techniques that were introduced since the first designs appeared (Rascon & Meza, 2017). This overview is necessary, as understanding the magnitude of different contributions can be counterintuitive.

2.1 Robotic Sound Source Localisation

As one can imagine, the first methods introduced for robotic SSL looked at natural systems and provided the basis of modern spatial localisation techniques (Lyon, 1983). Firstly, engineers around the globe developed efficient methods for representing spatial cues. After a couple of years they understood the limitations of their initial approaches, as some of their assumptions did not hold in more dynamic, common environments (Berglund & Sitte, 2005; Besson *et al.*,

2011). Researchers then started searching for different approaches and, as it is often the case, natural systems provided powerful metaphors that translated into the creation of more effective systems. More specifically, neuroscientific theories about SSL in animals opened the doors to a large family of bioinspired methods (Liu & Meng, 2007). In the following subsections we will travel from the initial systems using fixed microphone arrays to the most recent binaural biomimetic approaches.

2.1.1 First Generation: Static Microphone Arrays

Several approaches were taken during the 1990's to perform sound source localisation. Two spatial cues used since the first approaches are the Time-Difference-Of-Arrival (TDOA) between two or more microphones, and the variation of sound intensity or sound pressure level (SPL). As computing power was relatively scarce during this time, some implementations were optimised at the hardware level. In this way, Bhadkamkar (1994) designed customised hardware micro-components, to detect the TDOA between two microphones with a known *interaural* distance. The system of Bhadkamkar's CMOS chip for sound localisation utilises the TDOA between both microphones and can perform accurate SSL using low-frequency components of sound. However, the system does not compute SPL differences and is not able to localise sound sources when using high-frequency components that are part of the human audible range.

Another perspective could involve the integration of visual and auditory signals to disambiguate simultaneous sound sources (Nakadai *et al.*, 2000; Siracusa *et al.*, 2003; Nakadai *et al.*, 2010; Nakamura *et al.*, 2011). Interestingly, this approach was considered already in the mid-1990's. The system devised by Irie (1995) is an example of an early attempt to achieve multimodal sound localisation. He intended to support the localisation of sound sources in unconstrained environments with visual information. For this purpose, he implemented a feed-forward multi-layer perceptron. Unfortunately, the available computing resources at the time only made possible the classification of sound sources in three categories: left, right and centre. An interesting part of this implementation is that the network output has to be exactly zero to localise sources in front of the robot.

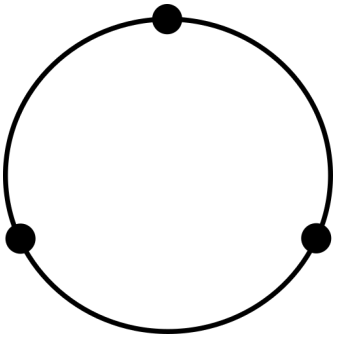


Figure 2.1: Array of 3 microphones in a ring. Array proposed by Huang *et al.* (1997a).

Hence, considerably lowering the localisation accuracy of sounds coming from the centre.

Huang *et al.* (1995) implemented zero-crossing algorithms to detect the sound source angle of incidence (Huang *et al.*, 1999). This method allowed him to estimate the difference in TDOA between three microphones in a ring (See figure 2.1). The system showed an excellent localisation performance for sounds coming from 360 degrees around the robot. However, the system importantly relied on the detection of sound onsets and was only tested in an anechoic chamber. Later on, they included an echo-estimation algorithm that facilitated the deployment of the system in reverberant environments (Huang *et al.*, 1997b,a). The system could satisfactorily detect the location of pure tones and claps. Onset detection is a promising approach to SSL (Newton & Smith, 2011), although a drawback from this approach was its poor performance for the detection of speech, as the onset of each frequency component dramatically varies. Finally, Huang *et al.* (1997a) successfully implemented a robotic system capable of detecting the spatial location of two concurrent speech signals in both, anechoic and reverberant rooms. A notable constraint of this system is the inability of dealing with frequency components above 2520 Hz. As a point of reference for the reader, the human audition can cope with frequencies up to 20000 Hz.

In order to increase the confidence of the TDOA estimations, researchers started increasing the number of microphones. Guentchev & Weng (1998) presented another kind of microphone array consisting of four sensors distributed

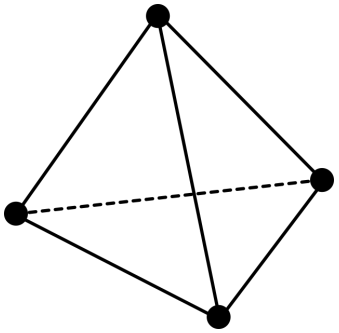


Figure 2.2: Array of 4 microphones in a pyramidal structure. Array proposed by Guentchev & Weng (1998).

in a pyramid-like structure (See figure 2.2). This system is very accurate and can perform 3D localisation, i.e., it can also estimate the distance to the sound source. It performs with an angle estimation error of ± 3° and a distance estimation error of ± 20%. Asono *et al.* (1999) implemented a near-field microphone array to localise sounds closer than 2 meters. The array consists of 8 microphones equally spaced in a ring. The main idea was to use information about the spatial location of a speaker to increase the Signal-to-Noise-Ratio (SNR) of the speech. The testing sounds included reverberation and an SNR of 20 dB. The authors tested the accuracy of the system with an automated speech recognition system using a vocabulary consisting of 492 words. With this system, it was possible to localise speech signals with an accuracy of 95-99%. The accuracy rate of the speech recognition system varied between 62-73%. As the sound localisation system relied only on TDOAs, the authors did not test it with frequency components higher than 3000 Hz, although the fundamental frequencies of human voice range between 60-7000 Hz.

The algorithms described so far have different weaknesses:

- They could not cope with SNRs lower than 20 dB, whereas natural systems can perform well with as low as 1 dB SNR (Guentchev & Weng, 1998).
- The presence of multiple sound sources would affect tracking any of them.
- Moving sounds were indistinguishable from a wider sound source.

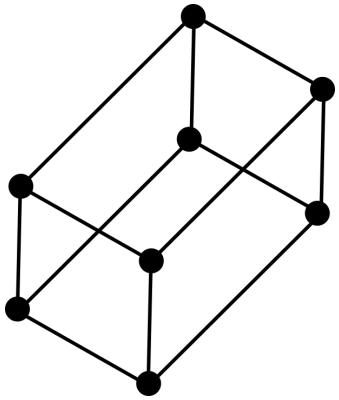


Figure 2.3: 8 microphones in a cubic array. Array proposed by Valin *et al.* (2003).

- The spectral content of the sound source could be a problem, as "sounds produced with a wide open mouth would yield a higher error value".
- It was difficult for the systems to perform well in places different to the environment in which the authors trained them.
- The absolute distance from the microphones to the sound source was a limitation, as 5 to 10 meters would already pose a serious problem.

As these problems are not present in natural systems, what can we learn from the physiological findings in animals? In the following subsection, we provide an overview of artificial SSL systems based on theories of sound localisation in humans, cats and guinea pigs.

2.1.2 Second Generation: Robotic Microphone Arrays

The systems described in subsection 2.1.1 achieved reasonably high accuracy for the localisation of sounds using the lower frequencies in the audible spectrum. Some of them were capable of performing accurately in partially reverberant environments, performing 3D sound source localisation or even localising two sources simultaneously. Those systems performed well in constrained environments, and even though such constraints varied among different approaches, none of them was capable of performing in diverse daily-life scenarios. For an SSL system

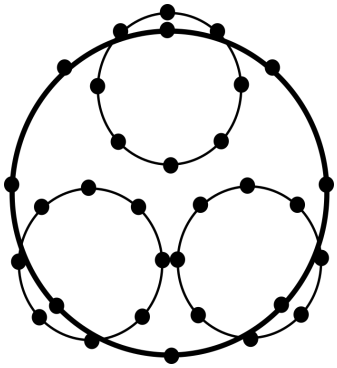


Figure 2.4: 32 microphones in a 4 rings array. Array proposed by Tamai *et al.* (2005).

to be reliable, it should be capable of handling SNRs present in everyday environments, reverberation, dynamic sources and simultaneous speakers (Hu *et al.*, 2006; Sasaki *et al.*, 2012). What was missing? Where did researchers find a need for improvement?

The available computational power continued growing exponentially and about a decade after the initial trials SSL systems adopted more sophisticated methods and increased the number of microphones. Valin *et al.* (2003) explored the performance of new spectral methods using an array of 8 microphones (See figure 2.3). The system could perform with an angular precision of 3° in the horizontal and vertical plane. In simulations, the array showed to be capable of estimating accurately the distance of a sound source up to 2 m away. Concerning the number of concurrent sources, the system could track only one source at a time. Tamai *et al.* (2005) designed an array of 32 microphones that could perform 3D SSL and the separation of simultaneous sound sources (See figure 2.4). They perform SSL with the delay and sum beamforming (DSBF) method and, in the following step, sound separation by integrating the DSBF method and frequency band selection (FBS). In this approach, the accuracy of the system reached up to 5° on the azimuth and elevation. The system can estimate the sound source distance with an error of less than 300 mm, but only when sound sources were closer than 1 m. This system can separate frequencies below 3300 Hz even when background noise is present.

High frequencies also contain useful spatial information and can improve sound source localisation and sound separation. However, none of the approaches using large microphone arrays takes advantage of the level differences produced by the shadowing of a head-like structure (Geng *et al.*, 2008; Cobos *et al.*, 2011; Nunes *et al.*, 2014). Here is where bioinspired approaches can offer guidance for integrating the information of such sound frequencies to develop more robust systems. In the following subsection, we introduce the advantages of bioinspired approaches by comparing some of the most representative methods.

2.1.3 Third Generation: Bioinspired Computation

The following biologically-inspired algorithms for sound source localisation and separation aim to apply neurophysiological theories to robotic systems. None of the described approaches pursues a complete emulation of the mammalian auditory pathway, as such a system would demand an amount of parallel computation that is not available in current hardware. Nevertheless, some natural principles have proven to be valuable paradigms for artificial sound source localisation (Agnes *et al.*, 2012; Amari, 2013; Chan *et al.*, 2010, 2012; Choudhary *et al.*, 2012). Artificial spiking neural networks (Maass, 1997) are of special interest for us, as this class of models share a *common language* that facilitates the representation of time-dependent information and its integration with additional sensory modalities (Maeder *et al.*, 2001; Karmarkar & Buonomano, 2007). Such common language between modalities is a fundamental property to create autonomous robots, as rich representations of the environment are essential to navigate in the real world (Hafting *et al.*, 2005; McNaughton *et al.*, 2006; Milford *et al.*, 2004; Milford & Wyeth, 2009).

Voutsas & Adamy (2007) created a model with multiple delay-lines using artificial spiking neural networks (Maass & Bishop, 2001; Maass *et al.*, 2002). After decomposing the sound in a set of fundamental frequencies, different delay values added to the sound waves allowed the estimation of ITDs. Their system only takes into account the ITDs and can localise broadband, and low-frequency sounds with 30° accuracy. However, the system performance decreases significantly for sounds with high fundamental frequencies, which is a common effect

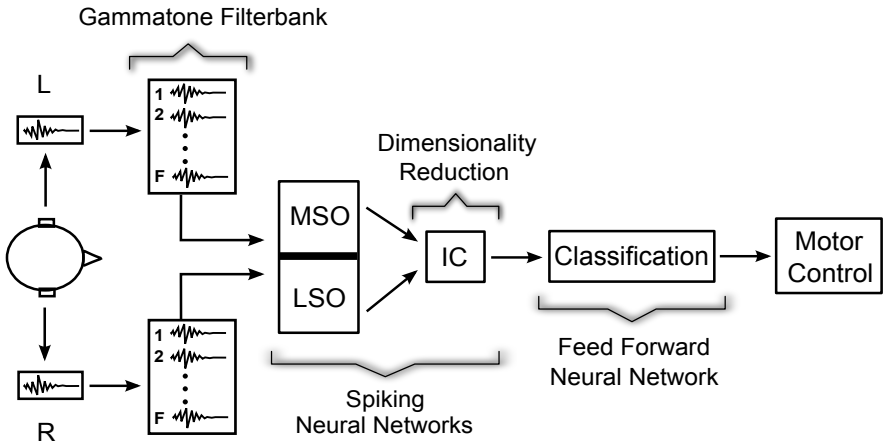


Figure 2.5: Sound source localisation architecture. Sound pre-processing consists of decomposing the sound input in several frequency components with the Gammatone filterbank emulating the human cochlea Slaney (1993). Afterwards, the MSO and LSO models *represent* ITDs and ILDs respectively. The IC model integrates output from the MSO and LSO while performing *dimensionality-reduction*. Finally, the *classification* layer produces an output angle that directs motor control (Rokni & Sompolinsky, 2012).

in systems relying only on temporal cues. The accuracy for localisation remains high with broadband signals, so their system performs better when it integrates information across a broader range of frequencies.

It is also possible to represent spatial information with more than two cues (Heckmann *et al.*, 2006; Rodemann, 2010). Rodemann *et al.* (2006) developed a model based on ITDs, ILDs and interaural envelop difference (IED). It can localise sound sources with a resolution of 10°, that is, with three times finer granularity than the system in Voutsas & Adamy (2007) using only one spatial cue. Nevertheless, the model in Rodemann *et al.* (2006) shows high sensitivity to the ego-noise produced by the robotic platform. The system computes the different localisation cues in parallel, and a weak winner-takes-all strategy defines the integration of the different cues. In all the testing conditions, higher frequencies lead to higher error rates estimating the sound source angle. A possibility for improvement could be to merge spatial cues with a non-linear model, as in the IC.

The systems from Willert *et al.* (2006) and Nix & Hohmann (2006) include

probabilistic models of the MSO, the LSO and the IC that can perform SSL with a resolution of 15°. In both cases, Bayesian statistics were used to estimate the connections between the layers and the systems perform robustly for simulated sound sources in real environments. A possible extension of this research is their implementation with ASNNs in order to explore the dynamics of neural populations and to exploit their robustness against noise (Ma *et al.*, 2006). Nevertheless, the results from these studies provide valuable insights precisely for the design of such biomimetic systems. Only Willert *et al.* (2006) mention multi-source tracking as part of their future work.

Murray *et al.* (2009) proposed an algorithm that relies mainly on the TDOA between a pair of microphones. He extracts the TDOA with a cross-correlation of both signals (Murray *et al.*, 2004). Afterwards, a recurrent neural network was capable of predicting the dynamics of the movement of a speaker. This approach demonstrates the benefits of motion prediction for continuous sound source localisation. The implementation of a head related transfer function (HRTF) was part of the future work for this project and would allow for SSL on the azimuth plane (Hornstein *et al.*, 2006; Keyrouz & Saleh, 2007).

Liu *et al.* (2010) proposes a biomimetic supervised learning algorithm for binaural SSL, where the MSO, LSO and IC are modelled using ASNNs and the connection weights are calculated using Bayesian inference (Futagi & Kitano, 2012). This system performs SSL with a resolution of 30° under reverberant and low noise conditions, and can also be used to track multiple moving sources. Dávila-Chacón *et al.* (2012) adapt the approach of Liu *et al.* (2010) to the Nao robotic platform (Gouaillier *et al.*, 2009) that produces ∼40dB of ego-noise. This neural model is capable of handling such levels of ego-noise and even increases the resolution of SSL to 15°.

In more recent work, Davila-Chacon *et al.* (2013) compare several neural and statistical methods for the *representation*, *dimensionality-reduction*, *clustering* and *classification* of auditory spatial cues. The evaluation of these neural and statistical methods follows a trade-off between computational performance, training time and suitability for life-long learning. However, the results of this comparison show that simpler architectures achieve the same accuracy as architectures

with an additional clustering layer. Figure 2.5 shows an overview of the best-performing SSL architecture. Davila-Chacon *et al.* (2013) found that a neural classifier on the top layer of our architecture is important to increase the robustness of the system against reverberation and ∼60dB of ego-noise produced by the humanoid iCub (Beira *et al.*, 2006). For this purpose, they include a feedforward neural network to handle the remaining non-linearities in the output from the IC model. Finally, in order to improve the robustness of the system to data outliers, they extended the architecture with softmax layers on the output of the IC model and the final layer of the SSL architecture.

More recently, research groups have developed novel SSL systems that can perform robustly under a variety of noise and reverberation Liu & Shen (2010a); Ren & Zou (2012); Pavlidi *et al.* (2013). The architecture introduced in Pavlidi *et al.* (2013) is particularly interesting, as it can estimate the number of sound sources present in the environment. Part of their suggested future work includes an adaptive width for the window analysing the input signals, as counting sound sources at low signal-to-noise ratio (SNR) requires different parameters than at high SNR. As a downside, these systems also neglect the spatial information encoded in high frequencies of sound sources. In the following section we introduce the biomimetic approach of Liu & Shen (2010a) and then describe the evolution of our computational model; from the simplifications to the spiking neural networks and the Bayesian model, to the extension of the model with additional neural and statistical layers.

2.2 Biomimetic Computational Model

This section describes in full detail our final biomimetic sound source localisation architecture. It has been designed from an embodied embedded cognition perspective to take advantage of the embodiment of the humanoid platforms used to test it. This approach reduces computational costs by using the embodiment of the robot as a passive sound filter, and helps to define the value of hyper-parameters in our models. For example, the biomimetic foundation constrains the topology of the connections between layers in our architecture (Oliver *et al.*, 2003).

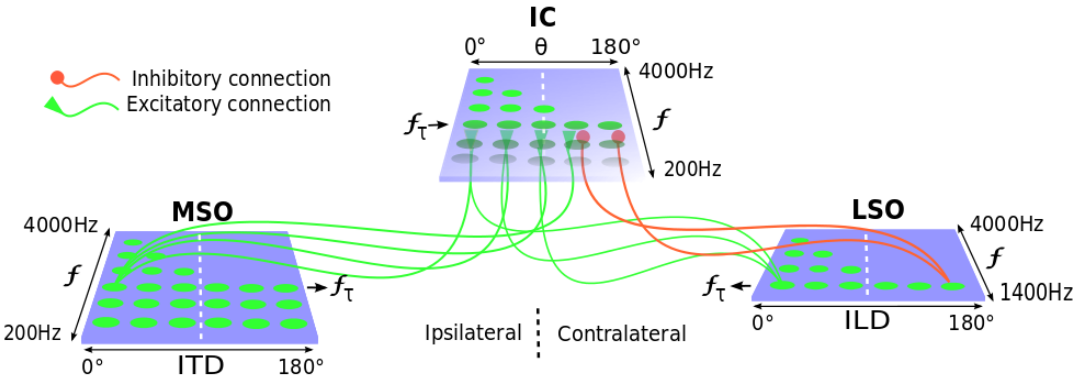


Figure 2.6: Topology of the connections between the MSO and LSO models to the IC model. The MSO has excitatory connections to the IC in f between 200 Hz and 4000 Hz, whereas the LSO has excitatory and inhibitory connections to the IC only in $f \geq f_\tau$ between 1400 Hz and 4000 Hz.

2.2.1 Cochlea Model

The first stage of our SSL architecture, shown in Figure 2.5, consists of a Gammatone filterbank modelling the frequency decomposition performed by the human cochlea Slaney (1993). This is, the signals produced by the microphones in the robot's ears are decomposed in a set of frequency components $f_i \in F = \{f_1, f_2, \ldots, f_I\}$. All the subsequent layers in our SSL architecture preserve the same tonotopic arrangement. In healthy young people, all consecutive f_i are logarithmically separated and respond to frequencies between ∼20 Hz and ∼20000 Hz Middlebrooks & Green (1991). We are primarily concerned with the localisation of speech signals; therefore we constrain the elements in F to the frequencies containing where most speech harmonics, i.e., between 200 Hz and 4000 Hz. Once the system decomposes both signals into I components, each wave of frequency f_i is used to generate spikes mimicking the phase-locking mechanism of the Organ of Corti, i.e., the model produces a spike when the positive side of the wave reaches its maximal amplitude.

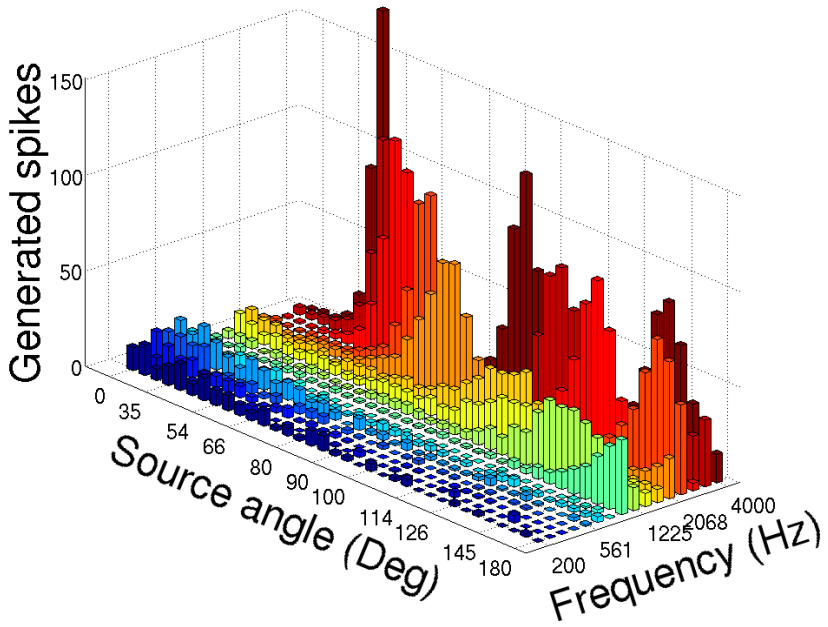


Figure 2.7: Activation of the **MSO model** for a sound consisting of white noise, presented to the robot at 15°. Notice that *lower frequencies* (blue) are more informative, as they produce a larger concentration of neural firing in the neurones sensitive to sounds produced around 15°, which is the real sound source angle, whereas *higher frequencies* (red) trigger the firing of neurones sensitive to sounds produced at the wrong angles.

2.2.2 Medial Superior Olive Model

Figure 2.6 depicts the biomimetic computational model that we designed following the neuroanatomy of the connections between the MSO and LSO layers to the IC layer. The MSO has excitatory connections to the IC in f between 200 Hz and 4000 Hz, whereas the LSO has excitatory and inhibitory connections to the IC only in $f \geq f_\tau$ between 1400 Hz and 4000 Hz.

In the following layer of the SSL architecture, we model the MSO as a mechanism to represent ITDs. As depicted in Figure 1.8, the computational principle observed in the MSO is modelled as a Jeffress coincidence detector Jeffress (1948) for each f_i. The MSO model has $m_j \in M = \{m_1, m_2, \ldots, m_J\}$ neurones for each f_i. The robot's interaural distance and the audio sampling rate constrains the value of m_J. Each neurone $m_{i,j} \in \mathbb{N}^0$ is maximally sensitive to sounds produced

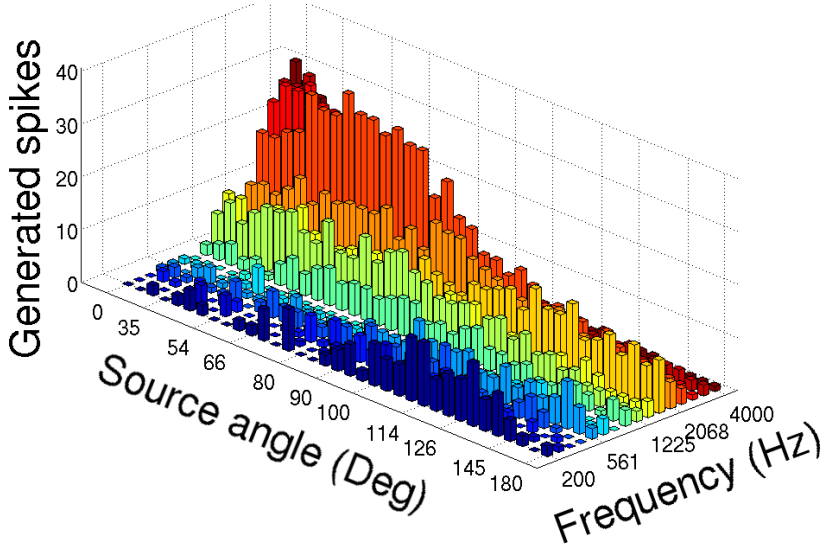


Figure 2.8: Activation of the **LSO model** for a sound consisting of white noise, presented to the robot at 15°. Notice that *higher frequencies* (red) are more informative, as they produce a larger concentration of neural firing in the neurones sensitive to sounds produced around 15°, which is the real sound source angle, whereas *lower frequencies* (blue) trigger the firing of neurones sensitive to sounds produced at the wrong angles.

at angle α_j. Therefore, $\mathbf{S}^{MSO}$ is the array of spikes produced by the MSO model for a given sound window of length ΔT. The mammalian auditory system relies mainly on delays smaller than half a period of each f_i for the localisation of sound sources (Schnupp *et al.*, 2011, Ch. 5.3.3). For this reason, the MSO model only computes ITDs when the time difference δt between two incoming spikes is smaller than half a period. This is, when $2f_i \cdot \delta t < 1$. Inspired by the mammalian neuroanatomy, the MSO model projects excitatory input to all $f_i \in F$ of the IC model (Meddis *et al.*, 2010, Ch. 4, 6.).

2.2.3 Lateral Superior Olive Model

At the same level of the SSL architecture, the LSO model represents ILDs. The system computes level differences by comparing the L and R waves from each f_i at the same points in time used for computing ITDs. The auditory system is known to compare the timing of neural spikes when the time delay between them

is less than half a period (Schnupp *et al.*, 2011, Ch. 5.3.3). Therefore, our MSO model considers the time difference Δt between t_1 and t_2 for the computation of ITDs, but not the Δt between t_2 and t_3. In order to determine the neurone that will fire, the LSO model computes ILDs as the logarithmic ratio of the vibration amplitudes at t_1 and t_2 as $\log(A_1/A_2)$ at times t_1 and t_2. The LSO model has $l_j \in L = \{l_1, l_2, \ldots, l_J\}$ neurones for each f_i. As the bit-depth of the sound data limits the value of l_J, it is possible to have many more neurones in the LSO than in the MSO. For the sake of simplicity, we chose to have the same number of neurones in the MSO and LSO models by setting $l_J = m_J$. Each neurone $l_{i,j} \in \mathbb{N}^0$ is maximally sensitive to sounds produced at angle α_j. Therefore, $\mathbf{S}^{LSO}$ is the array of spikes produced by the MSO model for a given sound window of length ΔT. Also inspired by the mammalian neuroanatomy, the LSO model projects excitatory and inhibitory input only to the highest frequencies $f_i \in F \mid f_i \geq f_\tau$ of the IC model (Meddis *et al.*, 2010, Ch. 4, 6.).

2.2.4 Inferior Colliculus Model

Then we arrive at the layer modelling the IC, where ITDs and ILDs are integrated. Figure 2.6 shows the topology of the connections between the MSO and LSO models to the IC model. Bayesian classifiers allow the continuous update of probability estimations and are known to have good performance even under strong independence assumptions (Rao, 2004). Furthermore, Bayesian classifiers allow fast computation as they can extract information from large dimensional data in a single batch step. For this reason, we estimate the connection weights assigned to the excitatory and inhibitory output of the MSO and LSO layer using Bayesian inference Liu *et al.* (2010). The IC model has $c_k \in C = \{c_1, c_2, \ldots, c_K\}$ neurones for each f_i. Each neurone $c_{i,k} \in \mathbb{R}$ is maximally sensitive to sounds produced at angle $\theta_k \in \Theta_K = \{\theta_1, \theta_2, \ldots, \theta_K\}$, where K is the total number of angles around the robot where sounds were presented for training. $\mathbf{E}^{MSO}$ and $\mathbf{E}^{LSO}$ are the ipsilateral MSO and LSO excitatory connection weights to the IC, and $\mathbf{I}^{LSO}$ are the contralateral LSO inhibitory connection weights to the IC. Therefore, $\mathbf{S}^{IC}$ is the array of spikes produced by the IC model for a given sound window of length ΔT. More precisely, $\mathbf{S}^{IC}$ is computed as follows:

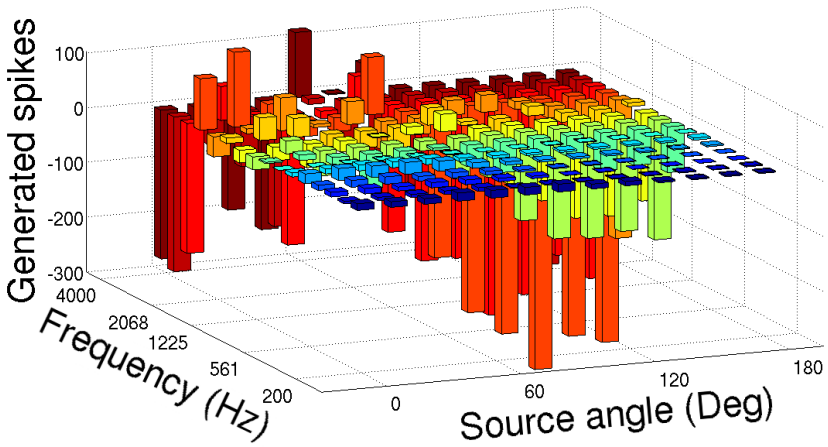


Figure 2.9: Activation of the **IC model** for a sound consisting of white noise, presented to the robot at 15°. Higher frequencies are represented in red and lower in blue. Notice that in comparison to the MSO and LSO models, the IC model has a more coherent spatial representation across *all frequencies* as a larger concentration of neural firing is found in the neurones sensitive to sounds produced around 15°, which is the real sound source angle. The IC model has fewer neurones than the MSO and LSO models to perform dimensionality reduction, and shows negative values as the inhibitory input is greater than the excitatory input from previous layers.

$$\mathbf{S}^{IC} = \mathbf{S}^{MSO} \odot \mathbf{E}^{MSO} + \mathbf{S}^{LSO} \odot \mathbf{E}^{LSO} - \mathbf{S}^{LSO} \odot \mathbf{I}^{LSO}. \tag{2.1}$$

Where $\odot$ indicates element-wise multiplication between the activation arrays. In order to estimate the connection weights $\mathbf{E}^{MSO}$, $\mathbf{E}^{LSO}$ and $\mathbf{I}^{LSO}$, we perform Bayesian inference on the spiking activity $\mathbf{S}^{MSO}$ and $\mathbf{S}^{LSO}$ for the known sound source angles Θ_K.

We define the set of training matrices obtained for each θ_k as $s_n \in S = \{s_1, s_2, \ldots, s_N\}$, where N is the total number of training instances. We describe first the Bayesian process used to estimate the connection weights between the MSO and the IC, where $s_n = \mathbf{S}_n^{MSO}$. Let $p(\mathbf{S}^{MSO}|\theta_k)$ be the likelihood that a sound that occurs at angle θ_k produces the spiking matrix $\mathbf{S}^{MSO}$. As we assume

Poisson-distributed noise in the activity of neurones $m_{i,j}$ in the MSO model,

$$p\left(\mathbf{S}^{MSO}|\theta_k\right) = \frac{\lambda_k^{\mathbf{S}^{MSO}} e^{-\lambda_k}}{\mathbf{S}^{MSO}!}, \ \forall k \in \Theta_K. \tag{2.2}$$

Where λ_k is a matrix containing the expected value and variance of each neurone m_{ij} in $\mathbf{S}^{MSO}$, and it is computed from the training set S for each θ_k. In a Poisson distribution, the maximum likelihood estimation of λ_k is equal to the sample mean, and we compute it as

$$\lambda_k = \frac{1}{N}\sum_{n=1}^{N} \mathbf{S}_n^{MSO}, \ \forall s_n \in S \mid \theta_k. \tag{2.3}$$

As we assume a uniform distribution over all angles in Θ_K, we assign the same prior $p\left(\theta_k\right) = 1/K$ to each θ_k. In order to normalise the probabilities to the interval $[0, 1]$, we compute the evidence $p(\mathbf{S}^{MSO})$ as:

$$p\left(\mathbf{S}^{MSO}\right) = \sum_{k=1}^{K} p\left(\mathbf{S}^{MSO}|\theta_k\right) p\left(\theta_k\right). \tag{2.4}$$

Afterwards, the posterior $p\left(\theta_k|\mathbf{S}^{MSO}\right)$ is computed using Bayes rule:

$$p\left(\theta_k|\mathbf{S}^{MSO}\right) = \frac{p\left(\mathbf{S}^{MSO}|\theta_k\right) p\left(\theta_k\right)}{p\left(\mathbf{S}^{MSO}\right)} = \mathbf{P}_k^{MSO}. \tag{2.5}$$

The same Bayesian inference process described so far is used for computing the LSO inhibitory and excitatory connections to the IC. Finally, the connection weights for each neurone $m_{i,j}$ in $\mathbf{P}_k^{MSO}$ and $l_{i,j}$ in $\mathbf{P}_k^{LSO}$ to neurone $c_{i,k}$ in the IC, are set according to the following functions:

$$\mathbf{E}^{MSO} = \begin{cases} \mathbf{P}_k^{MSO}, & \text{if } \mathbf{P}_k^{MSO} > \\ & \left(\omega_E^{MSO} . \arg\max_{\theta_k}\left(\mathbf{P}_k^{MSO}\right)\right), \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

$$\mathbf{E}^{LSO} = \begin{cases} \mathbf{P}_k^{LSO}, & \text{if } \mathbf{P}_k^{LSO} > \\ & \left(\omega_E^{LSO} . \arg\max_{\theta_k}\left(\mathbf{P}_k^{LSO}\right)\right) \\ & \bigwedge f_i \geq f_\tau \\ 0 & \text{otherwise} \end{cases}, \tag{2.7}$$

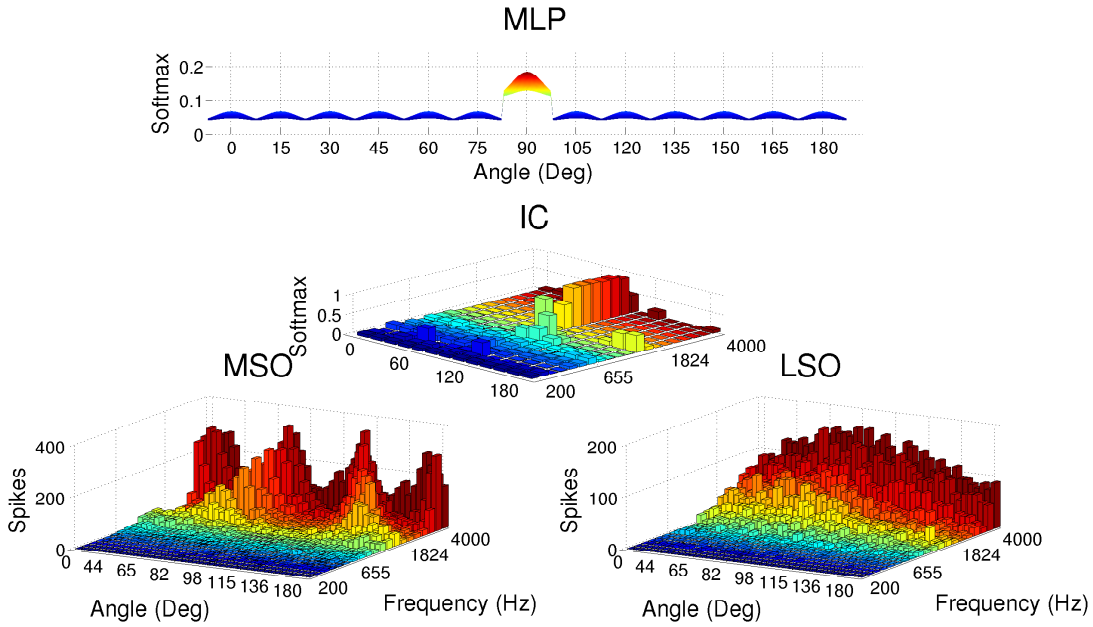


Figure 2.10: Output of all the layers in the SSL architecture for white noise presented in front of the robot (90°). Higher frequencies are represented in red and lower in blue. Notice that for this angle most of the IC frequency components agree on the sound source angle and the MLP correctly classifies the IC output.

$$\mathbf{I}^{LSO} = \begin{cases} 1 - \mathbf{P}_k^{LSO}, & \text{if } \mathbf{P}_k^{LSO} < \\ & \left(\omega_I^{LSO} . \arg\max_{\theta_k} \left(\mathbf{P}_k^{LSO}\right)\right) \\ & \bigwedge f_i \geq f_\tau \\ 0 & \text{otherwise} \end{cases} . \tag{2.8}$$

Where thresholds $\omega_E^{MSO} \bigwedge \omega_E^{LSO} \bigwedge \omega_I^{LSO} : \mathbb{R} \in [0, 1]$, determine which connections will be pruned. Following known neuroanatomy, such pruning avoids the interaction between neurones sensitive to distant angles (Liu *et al.*, 2008, 2009). The value of f_τ marks the transition between the lower and higher frequency spectrum. Figures 2.7, 2.8 and 2.9 show activation examples of the first version of MSO, LSO and IC models. This initial implementation did not assume Poisson-distributed noise in the activity of neurones, and it did not have the MLP and softmax layers described in subsection 2.2.5.

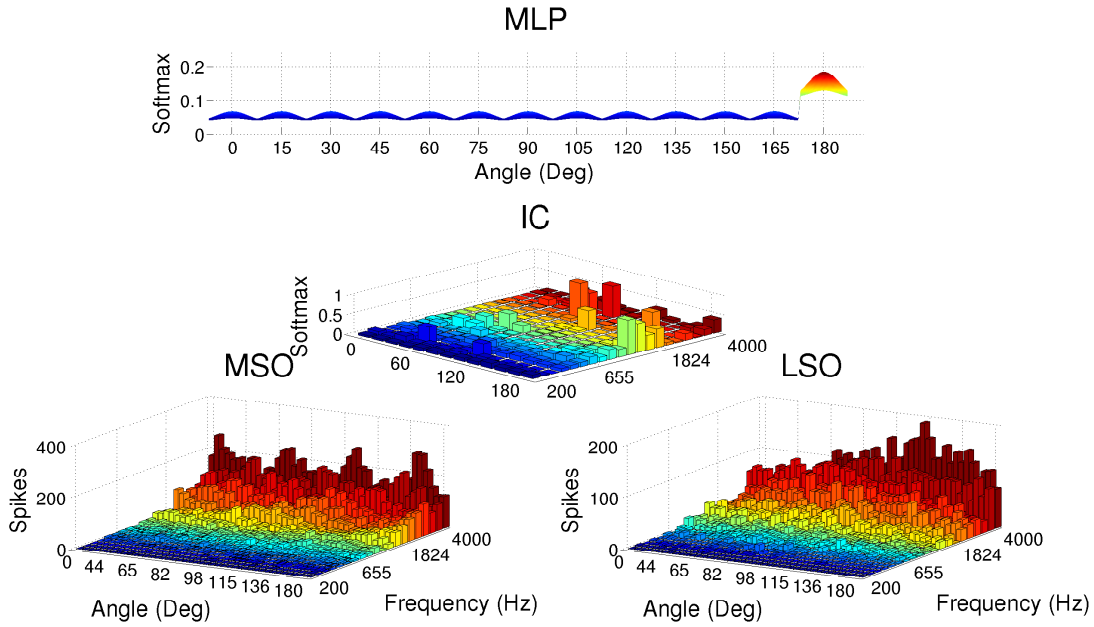


Figure 2.11: Output of all the layers in the SSL architecture for white noise presented on the right side of the robot (180°). Higher frequencies are represented in red and lower in blue. Notice that for this angle most of the IC frequency components disagree on the sound source angle; however, the MLP can cope with these non-linearities and correctly classifies the IC output.

2.2.5 Non-Linear Probabilistic Model

Finally, we use a feedforward neural network in the last layer of our SSL system for the classification of $\mathbf{S}^{IC}$. This layer increases the robustness of the system against ego-noise and reverberation. The output of the IC layer still shows non-linearities that reflect the complex interaction between the robot's embodiment and sound in the environment. Some of the elements that influence this interaction include the sound source angle relative to the robot's face, the head material and geometry, and intense levels of noise produced by the cooling system inside the robot's head. In previous work, we compare several neural and statistical methods Davila-Chacon *et al.* (2013) and found that a multilayer perceptron (MLP) was the most robust method for representing the non-linearities in $\mathbf{S}^{IC}$. The hidden layer of the MLP performs compression of its input as it has $|\mathbf{S}^{IC}|/2$ neurones, and similar to the IC neurones analysing a single f_i, the output layer of the MLP has

$c_k \in C$ neurones. In order to improve the robustness the system against data outliers, we perform softmax normalisation on $\mathbf{S}^{IC}$ before training the MLP:

$$\mathcal{S}^{IC} = \left(\frac{e^{\mathbf{S}_i^{IC}}}{\sum_{i'=1}^{I'} e^{\mathbf{S}_{i'}^{IC}}} \right), \ \forall f_i \in F, \tag{2.9}$$

and also on the output $\mathbf{S}^{MLP}$ of the MLP:

$$\mathcal{S}^{MLP} = \max_k \left(\frac{e^{\mathbf{S}_k^{MLP}}}{\sum_{k'=1}^{K'} e^{\mathbf{S}_{k'}^{MLP}}} \right), \ \forall c_k \in C. \tag{2.10}$$

Figure 2.10 shows the output of all layers in our SSL architecture after training it with a subset of utterances from the TIMIT speech dataset Garofolo *et al.* (1993). The figures show the spiking matrices produced by with white noise in order to depict more clearly the stereotypical patterns of each f_i. Notice that the hypotheses generated by most neurones in the IC layer agree on the sound source angle, irrespective of the frequency component f_i from which they receive input. In this case, it is not surprising that the MLP classifies correctly $\mathbf{S}^{IC}$ since a voting mechanism applying the *winner-takes-all* rule along each f_i would suffice for a correct classification. However, this is not always the case. Figure 2.11 shows an example of a more complex IC output. Notice that even when the hypothesis of most f_i in the IC layer disagrees, the MLP is capable of classifying correctly $\mathbf{S}^{IC}$.

2.3 Robotic Speech Recognition

The final step in this work is to explore the use of sound source localisation for improving the performance of Automatic Speech Recognition (ASR) in the context of robotic platforms (Weintraub, 1986; Sagi *et al.*, 2001; Rouat *et al.*, 2011). Here is where the biomimetic computation paradigm of this work meets with the embodied embedded cognition approach. As detailed in the following paragraphs, existing approaches can perform entirely accurate ASR with robotic platforms that support speech segregation with SSL (Maas *et al.*, 2011; Guo *et al.*, 2016; Zhang & Wang, 2017). However, there is still room for improvement, as often these methods are constrained by assumptions about the number of sound

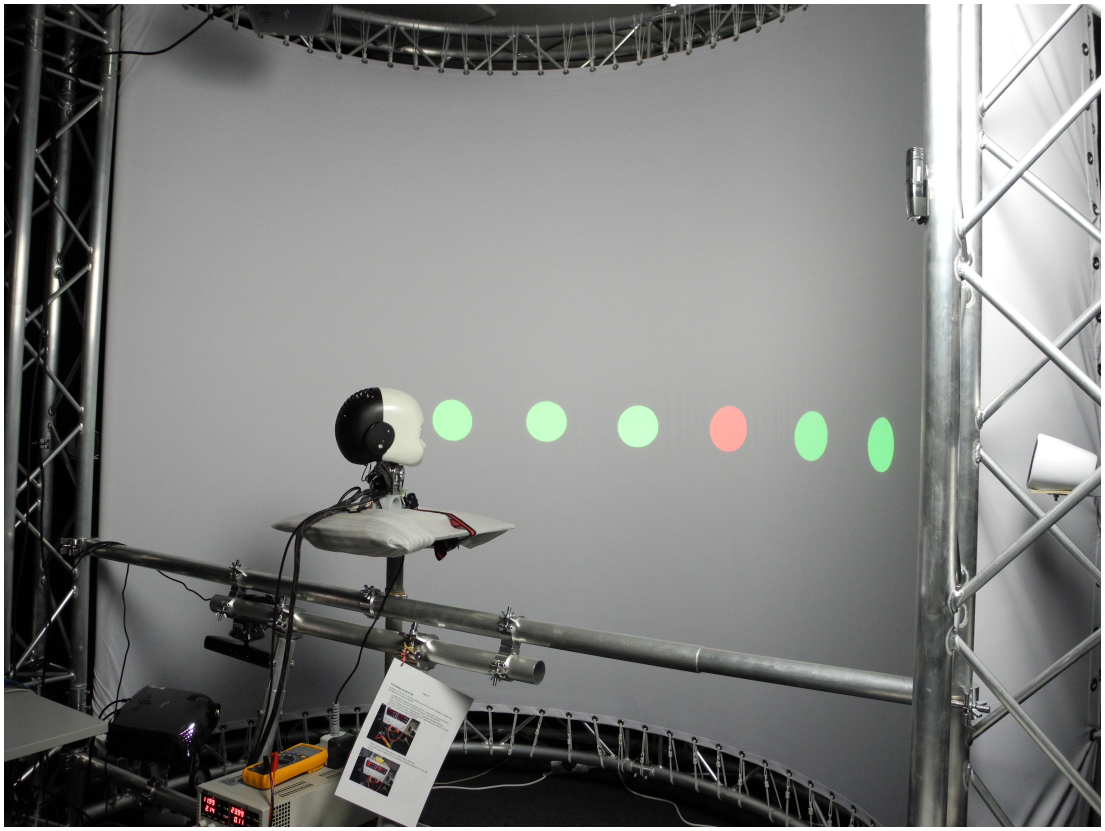


Figure 2.12: Experimental setup designed for research in cognitive robotics (Bauer *et al.*, 2012).

sources present in the environment, the amount of ego-noise and reverberation that they are capable of handling or the nature of the sounds that these systems are designed to localise (Ince *et al.*, 2011a,b; Wang *et al.*, 2018). All these systems have strengths that can complement each other and, hence, define the characteristics of the systems that we design afterwards.

Roman *et al.* (2003) present a system integrating SSL and ASR where speech recognition improves with the support of sound a source localisation system. They generated ITD-ILD binary masks that increased the SNR of the incoming speech signal, although the handling of moving speakers and reverberation remains open in their approach. Two other interesting examples in this direction are presented by Asano *et al.* (2001) and Fréchette *et al.* (2012). Both approaches make use of microphone arrays to localise speech sources in the environment. Afterwards,

they use information about the sound source to separate speech signals from noise in the background. The drawback of these methods is that they require prior knowledge about the presence and number of sound sources. Cong-qing *et al.* (2009) and Deleforge & Horaud (2012) present two alternative approaches that make use of binaural robotic platforms. However, both systems suffer from the same limitations of the binaural SSL methods discussed before as they mostly rely on the information contained in low-frequencies for SSL.

Woodruff & Wang (2013) present an architecture employing ITDs and ILDs for SSL and can perform segregation of an unknown number of sources. Nevertheless, the reported results consider at most two sound sources and segregation is performed offline due to the time required for computation. The approaches mentioned above rely on the construction of ideal binary masks for segregating speech from a discrete set of sound source angles. This approach presents an additional challenge because these methods are considerably affected when the sound source differs from the set of trained angles. Therefore, such approaches rely on an SSL system capable of tracking a human speaker almost instantly and with high accuracy. Our approach in the current work focuses on increasing the SNR of speech by continuously localising the most intense sound source and re-orienting the robot towards the speaker. In other words, we replace the use of ideal binary masks with a perception-action loop that maximises the SNR of sound arriving from the direction of the speaker. This sequential approach is feasible, given that our ASR system can recognise full sentences even if utterances have lower SNR at the beginning Twiefel *et al.* (2014); Heinrich & Wermter (2011a). In order to compare more clearly the performance of ASR with and without the support of SSL, we constrain the domain-independent output of an ASR system to a domain-dependent set of sentences. The experimental setup that we designed for research in multimodal integration for humanoid robots (Lim *et al.*, 2007), hence ideal for SSL and ASR, can be seen in Figure 2.12.

2.4 Conclusion

The development of SSL methods presented in this chapter provides a context for the development of our proposed method. It is interesting to see the increments in

the complexity of the hardware, going from arrays with a couple of microphones to arrays with several dozens of microphones and back to bioinspired binaural approaches. Also interesting, is the use of static platforms a few decades ago and the development of robotic systems that dynamically improve their accuracy by adapting their orientation to the sound source. Underlying this evolution, we find an increase in the available computing power and an improvement in the algorithmic approaches. Earlier systems made use of one spatial cue, whereas modern systems make use of multiple spatial cues that extract information from all the audible spectrum of sound. Finally, we approach systems that make use of the robot embodiment and use biomimetic computation that perform SSL more efficiently and is robust to noise and reverberation (Devore *et al.*, 2009). In the following chapters, we present the experiments that detail the evolution of our architecture and its application to the improvement of automatic speech recognition under high levels of ego-noise.

Chapter 3

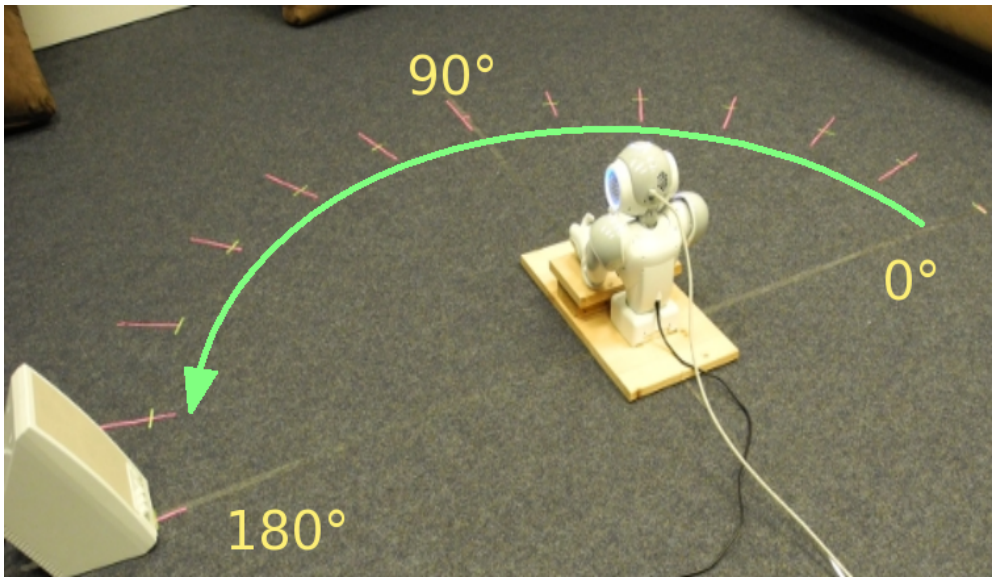
Noise-Robust Sound Source Localisation

In this chapter, we introduce the first version of a spiking neural network (SNN) used for binaural sound source localisation (SSL) integrated with a robotic platform with ego-noise. This SNN is based on the architecture developed by Liu *et al.* (2010) and has two main developments that differentiate it from the original model. The first improvement on the architecture is a simplification of the spiking model that replaces the *leaky integrate-and-fire* neurones (Stein, 1967), for linear inhibitory and excitatory neurones that always fire when stimulated. This simplification reduces the computational and memory cost, and interestingly, it improved the localisation accuracy. The second improvement is the determination of the maximum interaural level difference (ILD) produced by the geometry of the robot head used in the experiments. We determined the maximum ILD empirically by measuring the accuracy of sound source localisation while testing a range of different dB values.

3.1 Anechoic Room and Robot Nao

The experimental setup can be seen in figures 3.1a and 3.1b. The location was a room conditioned with heavy curtains that partially absorb the reverberation produced by the stimuli presented to the robot. The room reverberation time is $f \sim 0.4$ s and the sound pressure level $f \sim 25$ dB, what provides a recording

(a)



(b)

Figure 3.1: Sounds were played around the Nao in half circle ∅2 m, from 0° to 180° in 15° steps. 13 recordings were made in a room with reverberation damped by curtains.

environment quality close to the industry standards for music studios. In this room, we presented auditory stimuli from a semi-circle around the robot with a diameter of ∅2 m. The sound used for this experiment consists of white noise as it contains on average the same amount of energy across all frequency components. Along the semi-circle, we reproduced the stimuli from 13 positions between 0° and 180° in 15° steps. At this point in our research, we were interested in testing the SSL performance with standardised signals; therefore, we did not include more complex stimuli like human speech or *sum-of-ripples* (Klein *et al.*, 2000).

For the experiments detailed in the current chapter, we used the robot Nao (Gouaillier *et al.*, 2009). This robotic platform has been designed to help researchers in the field of humanoid robots, what fulfils our requirements of having a torso and a head-like structure between the ears. However, the head of the robot includes a fan for the cooling system that produces a background noise of 44.6 dBA at the right microphone and 41.6 dBA at the left microphone. In the original architecture of Liu *et al.* (2010), the experiments were carried out with a human-shaped head that did not produce internal or ego-noise; hence, our first objective is to find out if the SSL architecture is robust against the interference of high levels of stationary noise. According to specifications, Nao's distance between the left and right microphones is ∼0.12 m. Therefore, the highest frequency that does not generate interaural time difference (ITD) ambiguities is $f_\tau \approx 1400$ Hz.

3.2 Biomimetic Computation

As explained in Chapter 2.2, the spatial cues that we used for SSL were the ITDs and the ILDs. We extract these cues with the simplified models of the medial superior olive (MSO) and the lateral superior olive (LSO) and integrate their output with the Bayesian model of the inferior colliculus (IC). We estimate using Bayesian inference the weights of the connections from the MSO and LSO layers to the IC layer. However, in this first experiment we compute such connection weights with a simplified version of the model detailed in Section 2.2.4, as initially we do not include the assumption of Poisson-distributed noise in the activity of

neurones. Figure 3.2 details the resulting connectivity scheme between the MSO, LSO and IC models used in this first experiment.

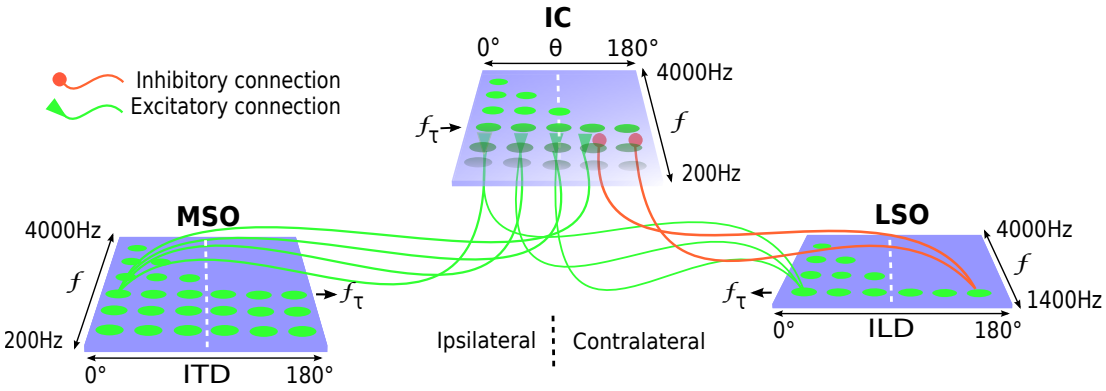


Figure 3.2: Multiple delay lines deliver spike-trains to MSO cells according to the Jeffress model (Schnupp *et al.*, 2011). MSO neurones respond to frequencies between 200 Hz and 4000 Hz. The difference of the wave amplitudes that produced a spike in the MSO is used to generate a spike in the LSO. LSO neurones respond to frequencies between ∼1000 and 4000 Hz. The MSO has excitatory connections to the IC in all frequencies. The LSO has excitatory and inhibitory connections to the IC in frequencies between ∼1000 Hz and 4000 Hz.

Willert *et al.* (2006) inspired this statistical inference, as the procedure is the same for computing the MSO excitatory connections, and the LSO excitatory and inhibitory connections. In the following paragraphs, we detail the method for estimating the MSO connections to the IC. First, we record one second of white noise with the left and right robot microphones at each of the θ_j^{IC} angles. A single loudspeaker produces the recordings at one meter from the robot. Afterwards, the recorded sounds are decomposed in n_f frequency components. The value of each frequency component f, for $f = 1...n_f$, is given by the gamma tone filter bank. Such filter simulates the mechanical filtering of the human cochlea. The MSO model analyses separately each of the n_f frequency components of sounds.

The IC model will have a total of n_{IC} neurones sensitive to each of the n_f frequency components. Let us define:

$$\theta_j^{IC} = \frac{180}{n_{IC} - 1} * j, \text{ for } j = 0, 1...n_{IC} - 1 \tag{3.1}$$

θ_j^{IC} is the angle of a sound source for which the IC neurone j, f is maximally sensitive, $S_{j,f}^{IC}$ is the number of spikes produced at the IC neurone j, f by a given sound and n_{IC} equals the total number of azimuth angles, in a half circle in front of the robot, where we place the sound sources for training.

The MSO model will have a total of n_{MSO} neurones sensitive each of the n_f frequency components. Let us define:

$$\theta_i^{MSO} = \frac{180}{n_{MSO} - 1} * i, \text{ for } i = 0, 1...n_{MSO} - 1 \tag{3.2}$$

θ_i^{MSO} is the angle of a sound source for which the MSO neurone i, f is maximally sensitive and $S_{i,f}^{MSO}$ is the number of spikes produced at the MSO neurone i, f by a given sound. The value of n_{MSO} depends mainly on the distance between the robot microphones, and the sample rate of the sound card used for recording the sounds defines its upper limit.

Now, lets define $p\left(S_{i,f}^{MSO}|\theta_j^{IC}, f\right)$, as the likelihood probability of a spike being produced at neurone i, f, given that a sound is being produced at angle θ_j^{IC}. $p\left(\theta_j^{IC}|f\right)$ represents the prior probability of a sound being produced at angle θ_j^{IC}, given that frequency component f is being analysed. Finally, $p\left(S_{i,f}^{MSO}|f\right)$ is the evidence of spikes being produced at neurone i, f.

Finally, from the previous definitions of likelihood, prior and evidence probabilities, the posterior probability can be computed with the Bayes rule:

$$p\left(\theta_j^{IC}|S_{i,f}^{MSO}, f\right) = \frac{p\left(S_{i,f}^{MSO}|\theta_j^{IC}, f\right) p\left(\theta_j^{IC}|f\right)}{p\left(S_{i,f}^{MSO}|f\right)} \tag{3.3}$$

Each of the n_{IC} training sounds recorded with the robot is analysed. First, the MSO model compares the wave from the left and right channels of a recording. It selects the first positive peak of the left channel wave as the *reference peak*. Then it selects as *comparison peaks*, all the positive peaks on the right channel located around one period after the reference peak. Depending on the time difference between the reference peak and each of the comparison peaks, the MSO model generates a spike in the corresponding MSO neurone i, f. Afterwards, the MSO model repeats the same procedure with the sides inverted and selects the first positive peak of the right channel wave as the reference peak. Finally, both reference peaks are shifted to the maximum peaks around one period further,

and both comparison processes repeated. The reference peaks stop being shifted forward when two periods are left before the sound wave finishes.

At the end of the analysis, an activation matrix A_j^{MSO} is generated for every training sound θ_j^{IC}. The system corrects the activation matrices by a proportion factor t_{prop}, relative to the size of the listening window that the robot will use under normal operation.

$$t_{prop} = \frac{t_{training}}{t_{listening}} \tag{3.4}$$

$t_{training}$ is the time length of the training sounds and $t_{listening}$ is the time length of the listening window that the system uses under normal operation.

Each activation matrix A_j^{MSO} is composed by n_f activation vectors $\overrightarrow{S_{j,f}^{MSO}}$. The connection weights are computed for every MSO neurone i, f in the activation vector $\overrightarrow{S_{j,f}^{MSO}}$, to the single IC neurone j, f. There are no connections between neurones sensitive to different frequencies. The connection weights are thresholded according to the following function:

$$E_{i,f}^{j,f} = \begin{cases} p\left(\theta_j^{IC}|S_{i,f}^{MSO}, f\right) & \text{if } p\left(\theta_j^{IC}|S_{i,f}^{MSO}, f\right) > \tau_{MSO} \max_f \left(p\left(\theta_j^{IC}|S_{i,f}^{MSO}, f\right)\right) \\ 0 & \text{otherwise} \end{cases} \tag{3.5}$$

$E_{i,f}^{j,f}$ is the excitatory connection from neurone i, f to neurone j, f and τ_{MSO} is the real number from the closed interval [01] that determines the minimum value of a posterior probability in order to be kept as a connection. A value of 0 represents no connection.

The likelihood probability is computed as follows:

$$p\left(S_{i,f}^{MSO}|\theta_j^{IC}, f\right) = \frac{S_{i,f}^{MSO}}{\sum_i S_{i,f}^{MSO}} \tag{3.6}$$

The system sums the spikes count in the activation matrix A_j^{MSO} over all the MSO neurones sensitive to the sound frequency component f.

The prior probability is the same for all neurones, as every angle θ_j^{IC} was trained once. Therefore, each angle has the same probability of producing a sound:

$$p\left(\theta_j^{IC}|f\right) = \frac{1}{n_{IC}} \tag{3.7}$$

n_{IC} equals the total number of sound source angles used during the training of the system.

The evidence probability is computed as follows:

$$\begin{aligned} p\left(S_{i,f}^{MSO}|f\right) = p\left(S_{i,f}^{MSO}|\theta_j^{IC}, f\right) p\left(\theta_j^{IC}|f\right) + \\ p\left(S_{i,f}^{MSO}|\neg\theta_j^{IC}, f\right) p\left(\neg\theta_j^{IC}|f\right) \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} p\left(S_{i,f}^{MSO}|\neg\theta_j^{IC}, f\right) p\left(\neg\theta_j^{IC}|f\right) = \\ \sum_{\neg j} \frac{S_{i,f}^{MSO}}{\sum_i S_{i,f}^{MSO}} \text{ for all } A_{\neg j}^{MSO} \end{aligned} \tag{3.9}$$

This simplification concludes the modifications made to the IC model from Liu *et al.* (2010).

3.2.1 Multi-Array Preliminary Study

In order to have a reference for the performance that is possible to achieve with the Nao platform, we performed a preliminary study using standard statistical methods for the extraction of time-difference-of-arrival (TDOA) between 3 microphone pairs in Nao's head (Li & Levinson, 2002). More specifically, we computed the cross-correlation between the signal of two microphone pairs: left-front and right-front. The system computed the TDOAs with a cross-correlation for a moving window, and the resulting TDOAs were concatenated to produce the input to a standard multilayer perceptron neural network (MLP). The MLP had $|I_I|= 2$ input neurones, was tested with $|I_H|= 6 \ldots 72$ hidden neurones, and had $|I_O|= 24$ output neurones. The network was trained with 4 speech recordings from 24 directions equally spaced around the robot.

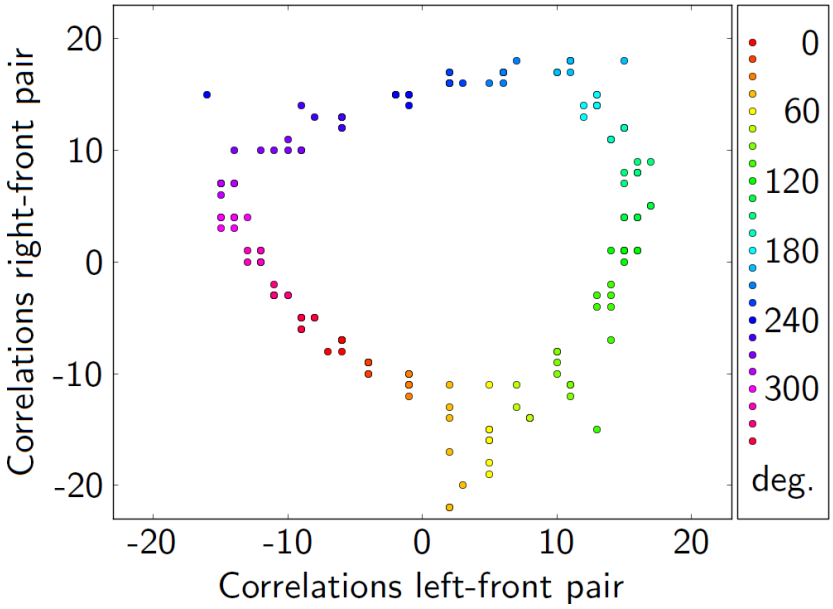


Figure 3.3: Cross correlation of ITD pairs.

3.2.2 Determination of Robot Interaural Level Difference

We hypothesise that the SSL architecture is robust enough to perform accurately under ∼40 dB of ego-noise. A crucial step to achieve this robustness is to determine the maximum ILD produced by the robot head. The maximum ILD resolution $\max_{ILD}$ that can be achieved by a robot depends on the geometry of its head, and it is necessary to estimate its value to extract ILDs from the high-frequency components of sounds. For SSL with humanoid dummy heads (Liu *et al.*, 2010) it is possible to determine the value of $\max_{ILD}$ from the scientific literature. However, for Nao's head, it was necessary to estimate the $\max_{ILD}$ by analysing the performance of the LSO model for SSL with different groups of frequencies.

3.2.3 Biomimetic Computation

We performed two experiments with the biomimetic SNN model. In the first one, we trained the robot with 1 s of uniform white noise (WN), and in the second one with a longer speech sequence. The speech sequence consisted of 4 instances

of the words *hello, look, fish, coffee* and *tea*, each pronounced by male and female speakers. We chose the words to contain tonal sound from vowels as well as fricative and plosive sound from consonants. In this way, we could observe the effect these phonological subclasses may have on the system accuracy if any. While performing SSL, the recordings made by the robot were split into 16 frequency components between 200-4000 Hz as shown in Figure 3.2. We chose the range of frequencies to contain most the harmonics produced in speech (Titze & Martin, 1998; Baken & Orlikoff, 2000) and determined the frequency components by applying the Patterson-Holdsworth filter bank algorithm (Holdsworth *et al.*, 1988; Slaney, 1993). In both experiments, the testing sounds consisted of instances of the same words not used during training and of 0.25 s samples of white noise.

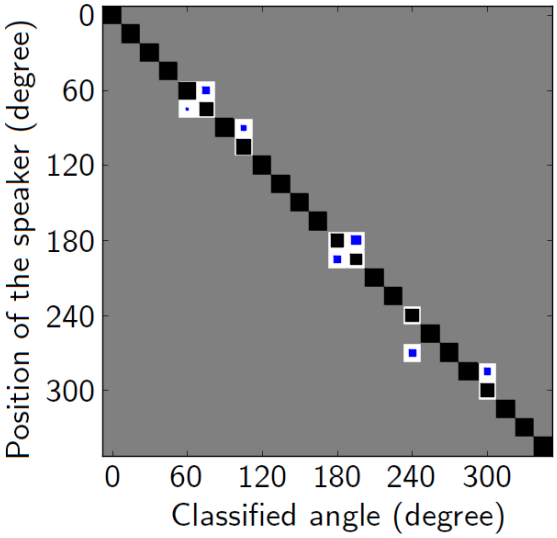


Figure 3.4: Confusion matrix of the MLP output.

3.3 Experimental Results

In order to get insights into the TDOAs produced between the two pairs of microphones, we computed the phase shifts producing the highest correlation for stimuli presented 360° around the robot. In this experiment, the extension to the full circle was straightforward as we were using 3 microphones. Figure 3.3 shows the results of the cross-correlations between the left and front microphones.

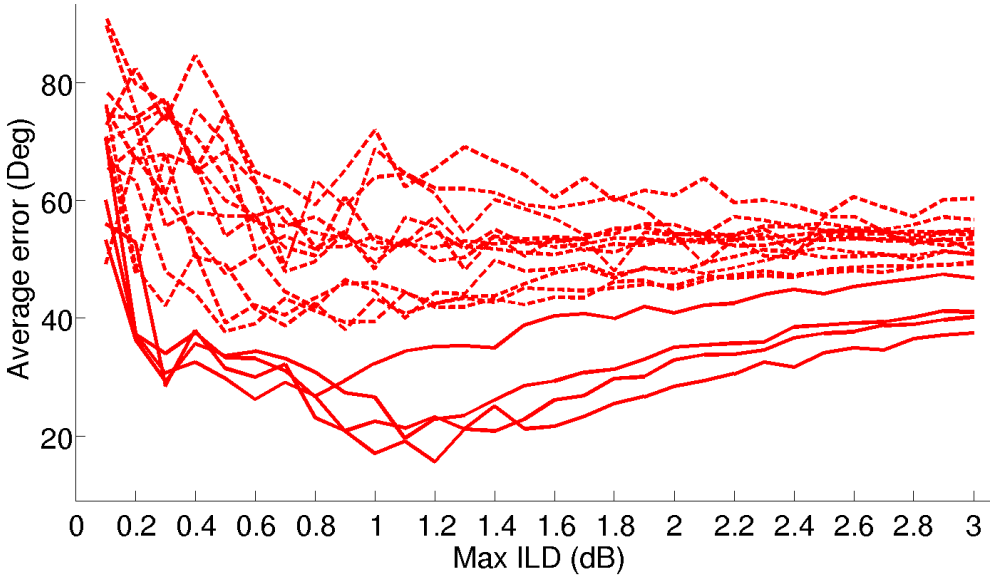


Figure 3.5: Localisation error of the LSO model when using the *winner-takes-all* strategy. Solid lines represent higher frequency components, and dotted lines represent lower frequency components. The average error for high-frequency components is lower than the average error for low-frequency components, which is consistent with the results obtained in neurophysiological experiments with animals. Even between high-frequency components, the average error tends to decrease around a small range of interaural level differences close to 1 dB. This shows that customisation of the $\max_{ILD}$ value can reduce by half the average localisation error, even between the best performing frequency components.

The results of this first experiment show that the system can perform multi-aural SSL using the robot Nao, independently of the ego-noise produced by the cooling system. As can be seen in Figure 3.4, the network classified the location of the sound source for most of the source angles correctly, and when the classification was erroneous, the magnitude of the error was minimum. Overall, the MLP classified the sound source angles with an accuracy of 91%.

Using more than two microphones was avoided in the following experiments for the sake of biological plausibility. Therefore, the question remained whether the binaural approach was going to perform as accurately as the approach with 3 microphones. The second experiment was related to the customisation of the $\max_{ILD}$. Once determined, the $\max_{ILD}$ value remained fixed in the following

experiments. We estimated the best LSO performance for all –and each– of the n_f frequency components from a range of $\max_{ILD}$ values between 0.1 dB and 3.0 dB at 0.1 dB steps. Figures 3.5 and 3.6 show the MSO and LSO output errors plotted against all the tested $\max_{ILD}$ values.

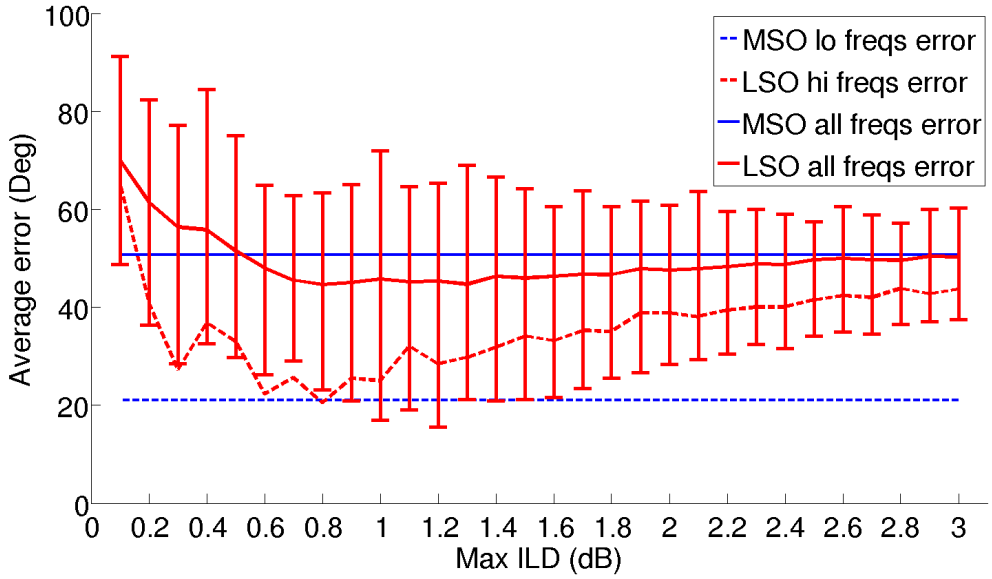
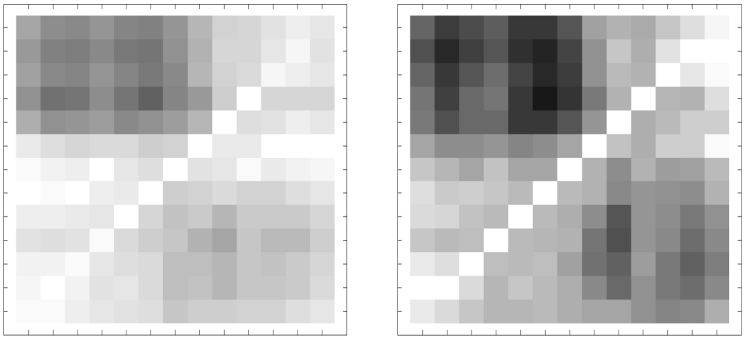


Figure 3.6: MSO and LSO output errors estimated from their best frequency components (dotted lines), and average error from all frequencies (solid lines). In both cases the MSO error (blue lines) is constant for all $\max_{ILD}$ values and the best performance was reached at ∼0.8 dB.

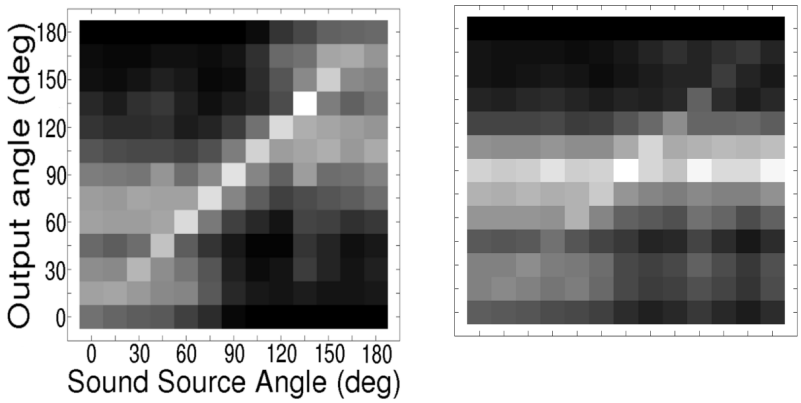
Finally, we needed to analyse in more detail the behaviour of the system with the $\max_{ILD}$ value that provided the best performance. The third experiment is related to SSL with only two microphones using a biomimetic spiking neural network. The results showed that the system was capable of differentiating sounds with a granularity of 15°, but with higher error rates than the array of three microphones. The confusion matrices can be seen in Figure 3.7. More specifically, the system has better performance when training with speech and testing with WN, but the accuracy diminishes when the system is trained with WN and tested with speech. These results are somehow unexpected, as training with WN has yielded good results in previous versions of the architecture.

Figures 3.8 and 3.9 compare the localisation error of the MSO, the LSO and



(a) Trained with S and tested with S.

(b) Trained with S and tested with WN.



(c) Trained with WN and tested with WN.

(d) Trained with WN and tested with S.

Figure 3.7: IC output confusion matrices when the system was trained or tested with uniform white noise (WN) or speech (S). The speech output is for the word *fish*. Lighter areas indicate higher values.

the IC models. In this case, the output of each model was chosen using an average of the winner-takes-all strategy applied to each frequency component. As expected, the system achieves the best localisation performance when tested with WN. Except for 0° and 180°, the IC lower boundaries in Figure 3.8 show no deviation from the ground truth values when localising WN stimuli. Figure 3.9 details further the output of the IC alone. In general, for all angles and all sound classes, the IC performance highly improves in comparison to the classification accuracy of the MSO and the LSO models alone as the localisation error drops

to zero between 60° and 120°.

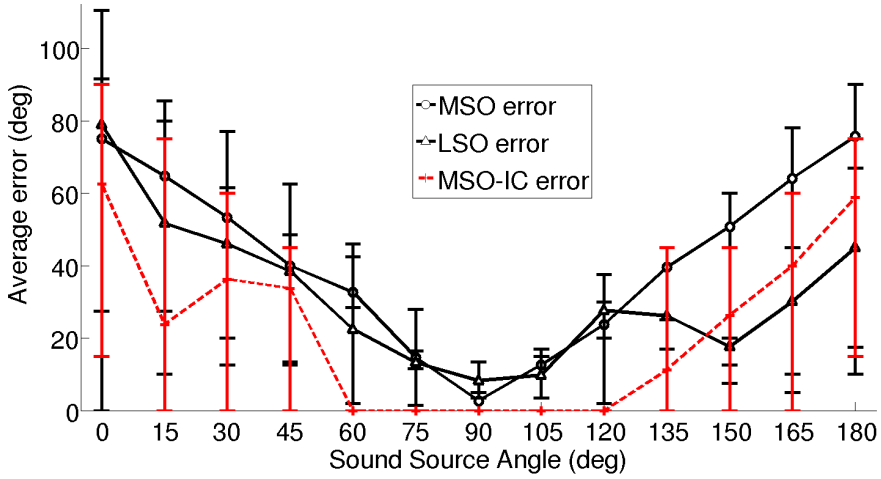


Figure 3.8: MSO, LSO and IC average errors when training with speech and testing with WN and speech. The errors are averaged over all sound classes. Notice that the IC has higher accuracy than the MSO and LSO.

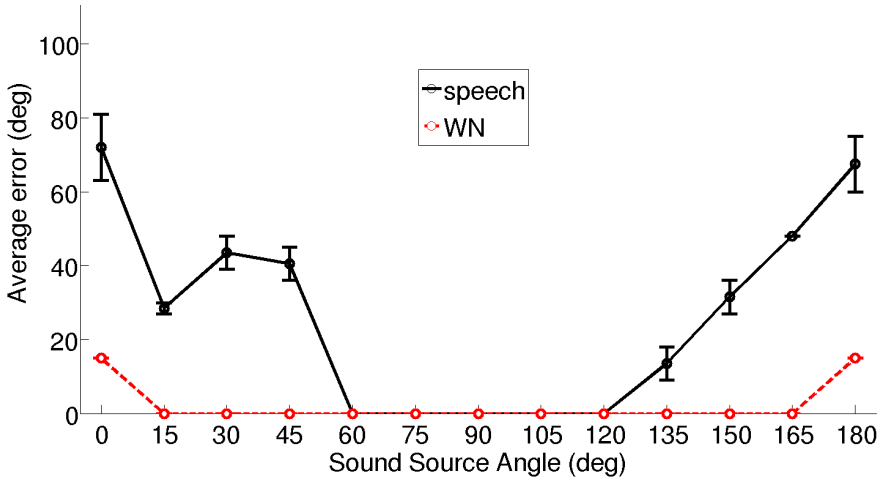


Figure 3.9: IC average errors when training with speech and testing with white noise or speech. Notice that the localisation error for white noise is zero for most angles. It is interesting to notice that training with speech produces a system that performs better than the common practice of training with white noise.

3.4 Conclusion

In this chapter, we confirmed the robustness of a biomimetic approach to SSL, even though the architecture consisting of 3 microphones performed more robustly. The integration of auditory cues in the IC showed higher performance than the MSO and the LSO alone. The IC model made no error in the 60° in front of the robot and had near perfect localisation accuracy for WN. The optimised algorithm proved to be capable of segregating sound sources with similar precision to state-of-the-art algorithms (Nix & Hohmann, 2006; Willert *et al.*, 2006; Voutsas & Adamy, 2007).

Estimating the optimal $\max_{ILD}$ value for Nao's head, allows the system to double the resolution for the localisation of sound sources in comparison to Liu et al. (Liu *et al.*, 2010), where the system performs SSL with a granularity of 30°. We found the $\max_{ILD}$ through an analysis of the LSO activation across all frequency components, but it was clear that high-frequency components are better suited for the extraction of ILDs. Furthermore, the frequency decomposition of the input signals opens the possibility of localising concurrent and dynamic sound sources (Liu *et al.*, 2010) that have different harmonic components. Such an advantage is missing in networks that extract ITD pairs from the cross-correlation of the sound wave.

The Bayesian inference process allowed the system to perform more robustly under high levels of ego-noise. When the MSO and LSO models were presented only with the robot's ego-noise, their output was a fixed angle. However, such ego-noise activation is distributed evenly among the IC neurones, and as expected their overall output cancels out. Equally important, the IC regularised the output of the system in a way that makes it more suitable to human interpretation and varies more linearly than the output of either the MSO and the LSO models. The IC output can be designed with a reduced number of neurones, making it useful as a dimensionality reduction model. This reduction helps to speed up the training of layers added in subsequent extensions of our architecture.

The processes underlying spatial hearing can be used for the segregation of speech and increase its signal-to-noise-ratio in this way (Roman *et al.*, 2003). In the following experiments, we go further in this direction, as we explore the

potential of using SSL for the enhancement of automatic speech recognition. Ultimately, we pursue a multimodal approach to the long-standing *Cocktail Party Problem*, and SSL is an essential ingredient in such enterprise (Even *et al.*, 2011; Li *et al.*, 2012; Kim *et al.*, 2015).

Chapter 4

Static Sound Source Localisation

When deploying binaural sound source localisation (SSL) algorithms in different environments and robotic platforms, it is crucial to use methods that are robust against diverse sources of noise and reverberation. In order to asses this challenge, in this chapter we compare the performance of various methods that could fulfil the same function at each stage of the SSL system that we propose. The architecture has three *degrees of freedom*, i.e. each tested architecture employs a different combination of *representation* of binaural cues, *clustering* and *classification* algorithms. The heuristic for the selection of methods is the same at each degree of freedom: to compare the impact of traditional statistical techniques versus machine learning algorithms with different degrees of biological inspiration. We evaluate the overall performance in the analysis of each system, including the accuracy of its output, training time and adequateness for life-long learning. The results support the use of hybrid systems, consisting of diverse kinds of artificial neural networks, as they present a practical compromise between the characteristics evaluated.

4.1 VR Room and Robot iCub

Figure 4.1 depicts our experimental setup. It consists of a humanoid robotic head immersed in a virtual reality (VR) setup designed for audio-visual integration (Bauer *et al.*, 2012). This setup can help tremendously to test neural architectures inspired in natural systems (Rucci *et al.*, 1997) with robotic platforms (Rucci

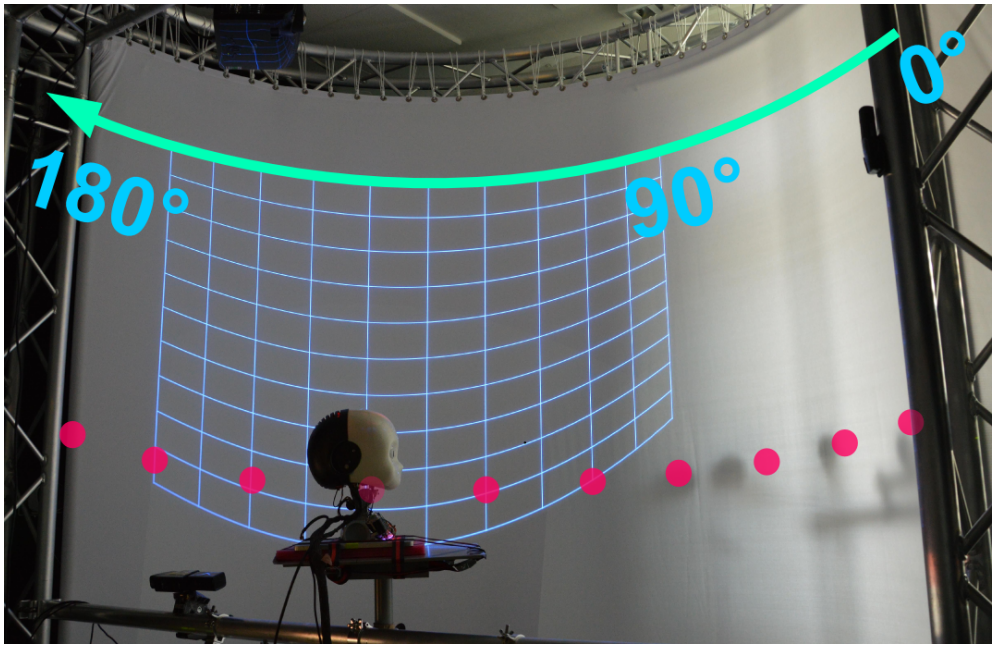


Figure 4.1: Audio-visual VR experimental setup. The grid shows the curvature of the projection screen surrounding the iCub humanoid robot head and the dots represent the location of the sound sources behind the screen.

et al., 2000). The iCub is a platform designed for studies in embodied cognition and cognitive developmental robotics (Beira *et al.*, 2006). The iCub head has a geometry similar to the average 3 to 4 years old child and is equipped with a pair of microphones surrounded by pinnae.

The position of the head remains fixed during the experiments, what we denominate *static* SSL. The iCub head produces ∼60 Hz of ego-noise, which is one of the most common challenges in humanoid platforms. In order to reduce the influence of additional variables, we reduce the reverberation in the room with damping curtains. The stimuli that we present to the robot consist of $0.25ms$ segments of white noise (WN) and the words *hello, look, fish, coffee* and *tea* recorded from male and female subjects. The WN class consisted of 12 instances, and the speech class consisted of 40 instances of each word. Each instance of both sound classes is presented once to the iCub between 0° and 180° at 15° steps along the azimuth plane. We present stimuli at the same elevation angle and a distance of $\sim 1.3m$.

The system is tested with four *training / testing* configurations: *Speech / Speech*, *WN / WN*, *WN / Speech* and *Speech / WN*. As expected, we obtained the highest performance when training and testing with different instances of the same class of sounds, i.e. with the *Speech / Speech*, *WN / WN* configurations. The lowest performance came from the *WN / Speech* configuration. However, some architectures were able to generalise between classes in the *Speech / WN* configuration. For this reason, we focus in this chapter on the results obtained with the *Speech / WN* configuration as it is interesting to analyse the generalisation achieved by the learning process.

4.2 Neural and Statistical Processing of Spatial Cues

We implement an architecture with three degrees of freedom in order to compare different SSL systems. The architecture is depicted in Figure 4.2. Each degree of freedom represents a layer, or processing step, that can be accomplished by alternative methods. The architecture layers consist of *preprocessing*, *representation*, *clustering* and *classification* of binaural sound input. Within these layers, the system performs the preprocessing step with a fixed algorithm; therefore, we do not consider it as a degree of freedom. During this step, the sound input is decomposed in several FCs with the Patterson-Holdsworth filter bank (PHFB) (Slaney, 1993).

The representation layer is in charge of characterising ITDs and ILDs numerically. The clustering layer is an intermediate step that can potentially improve the performance of classifiers, as it can distribute a large number of prototype vectors similarly to the underlying distribution of the training data. The clustering layer is not present in some of the tested systems, as it is also possible to directly classify the output of the representation layer. Finally, the classification layer generates an output angle that can we use for motor control (Rokni & Sompolinsky, 2012). In the following subsections, we detail further each of the processing layers in the architecture.

4.2.1 Preprocessing of Sound Signals

The first stage in our SSL system is the PHFB. This filter decomposes the left (L) and right (R) sound recordings in frequency components $f \in \{1, 2 \dots F\}$, where $F = 20$. The filterbank separates the extracted f on a logarithmic scale between 200 Hz and 4000 Hz, i.e., with an increase in bandwidth that resembles the response of the human cochlea. Afterwards, the subsequent layers compare the corresponding f from L-R signals for the extraction of spatial cues. All the classification methods that we describe in this chapter use this step (see Figure 4.2).

4.2.2 Representation of Spatial Cues

The basis of SSL algorithms is the set of localisation cues chosen as input. As the method used to represent spatial cues can influence the accuracy of the system's output, we want to compare the performance of our SSL system when representing spatial cues with traditional signal processing techniques against bioinspired methods. For this reason, we choose two of the most representative methods in binaural SSL research for representing ITDs: Cross-correlation (CCR) (Benesty *et al.*, 2007) and MSO Jeffress coincidence detector (Liu *et al.*, 2010; Joris *et al.*, 1998). We also make use of ILD cue and represent it with an LSO model previously presented by the authors (Liu *et al.*, 2010). Furthermore, we compare two integration methods for the MSO and LSO outputs. The first method (MSO-LSO) appends the output of the MSO and LSO models, and the second method (Bayes IC) integrates the output of both models using Bayesian inference. In Figure 3.2 are shown further details on the MSO, LSO and IC models. In the following sub-subsections we detail each of the representation methods.

4.2.2.1 Cross-Correlation

The Cross-Correlation (CCR) technique is used to estimate the cross-correlation sequence $CCR_{L,R}$ between L and R input signals, assuming them to be random

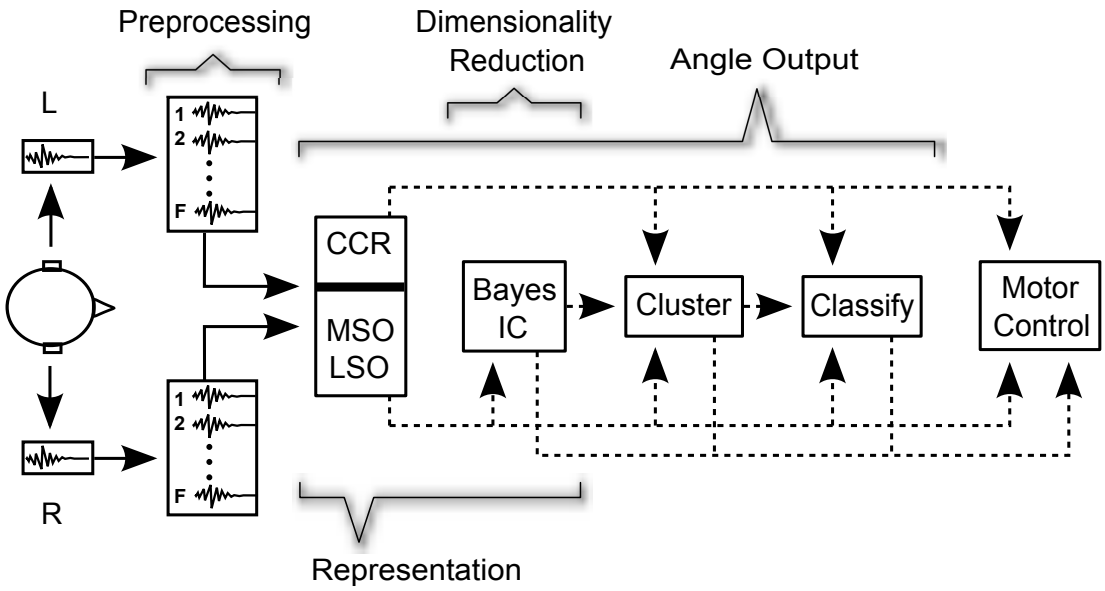


Figure 4.2: Testing architecture. Solid lines represent fixed steps, and dotted lines represent different systems that assembled for motor control. The shadowed brackets indicate the layers that the testing meta-architecture combines to define different architectures. The *preprocessing* consists on decomposing the sound input in several FCs with the Patterson-Holdsworth Filter Bank (PHFB) (Slaney, 1993). Then, the *representation* layer numerically characterises different spatial cues. Alternatively, the representation provided by the Bayes IC integrates output from the MSO and LSO in vectors with reduced dimensionality. The *clustering* layer distributes a larger amount of prototype vectors in the space of represented cues in order to test for a possible improvement in SSL accuracy. All systems were tested with and without this intermediate layer. Finally, the *classification* layer produces an output angle that can be used for motor control.

stationary processes sampled from time window Δt.

$$CCR_{L,R}(j, f, \Delta t) = \begin{cases} \sum_{i=0}^{J-1} L_{i,f,\Delta t} \cdot R_{i-j,f,\Delta t}, & \text{for} \quad 0 \leq j \leq J \\ CCR_{L,R}(-j, f, \Delta t), & \text{for} \ -J \leq j < 0 \end{cases}, \tag{4.1}$$

where i represents sampled values from the input signals, j are the ITD shifts made when computing the correlation sequence and J is the length of the input signals.

We use the correlation sequences of all f as input to the clustering or the classification algorithms. However, the output angle Θ can also be estimated directly from the j that maximises the correlation over all f with the *winner-takes-all* (WTA) rule:

$$ITD_{win} = \arg\max_{j} \left(\sum_{f} CCR_{L,R}(j, f, \Delta t) \right) . \tag{4.2}$$

We are interested in using WTA for benchmarking, as it is the classification technique the authors previously used in the MSO, LSO and IC models (Liu *et al.*, 2010; Dávila-Chacón *et al.*, 2012). Due to the geometry of the head, ITDs vary non-linearly as a sound source moves around us. Therefore, the output angle is computed as follows:

$$\Theta = \sin^{-1} \left(\frac{ITD_{win} - ITD_{max} + 1}{ITD_{max}} \right) , \tag{4.3}$$

where ITD_{max} is the maximum possible ITD that occurs when the sound source is aligned with the interaural axis.

4.2.2.2 Medial Superior Olive

One of the methods we use for extracting ITDs is Liu et al. (Liu *et al.*, 2010) SNN model of the MSO. This method takes inspiration from neurophysiological theories describing the underlying mechanisms of the MSO (Ashida & Carr, 2011), including the Jeffress Coincidence Detector model. After decomposing the sound signals with the PHFB, each frequency component f is phase-locked to its positive values. This locking means that hair cells in the organ of Corti reach the highest probability of producing a spike when the amplitude of vibrations in the basilar membrane is maximal (Richter *et al.*, 1998).

Afterwards, the system compares the maximum positive values in time window Δt, and the phase shift between these maximums is used to estimate the ITD. In the last step, neurones $k \in \{1, 2 \dots K\}$ in the MSO respond to different ITDs and for every time window Δt generate a spikes matrix $\mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}}$ of size $F \times K$.

We can feed the classification algorithms with $\mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}}$, or directly compute the output angle Θ from the k with maximal neural activity among all the f. For

the latter case, ITD_{win} could be estimated using the WTA rule as in eq. (4.4) and Θ as in eq. (4.3).

$$ITD_{win} = \arg\max_{k} \left(\sum_{f} \mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}} \right) . \tag{4.4}$$

4.2.2.3 Lateral Superior Olive

For estimating ILDs we use Liu et al. (Liu *et al.*, 2010) SNN model of the LSO. It also was developed by some of the authors and our current objective is to test it in a different anthropomorphic head with ego-noise. In the LSO model neurones $k \in \{1, 2 \dots K\}$ fire depending on differences in L-R amplitudes for each f. Using the same pairs of L-R values from which ITDs are measured, ILDs are computed as $\log(L_{f,t}/R_{f,t})$. Therefore, at every time step Δt a spikes matrix $\mathbf{S}^{\mathbf{LSO}}_{\mathbf{\Delta t}}$ of size $F \times K$ is generated. Afterwards, we obtain the output angles following the same procedure applied to $\mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}}$.

4.2.2.4 Inferior Colliculus

Reducing the dimensionality of input vectors can decrease the amount of data and time required for training machine learning algorithms. For this reason, we also test the clustering and classification algorithms with an integrated version of the MSO and LSO output vectors. Such integrated vectors are constructed using Bayesian inference in a model of the inferior colliculus (IC) (Liu *et al.*, 2010). A significant computational advantage comes from the IC dimensionality reduction, as IC output vectors are more than six times smaller than the MSO and LSO output vectors together. More details of the IC integration architecture are shown in Figure 3.2.

An additional benefit from this integration process comes from the overlap of MSO excitatory connections and LSO inhibitory connections. The LSO captures the useful information for SSL contained in high frequencies but generates ambiguous information from low-frequencies. The MSO captures the useful information for SSL contained in all frequencies, but also generates ambiguous information from high-frequencies. For this reason, LSO inhibitory connections

can help to remove misleading information generated by the MSO at high frequencies. Therefore, the IC provides a more accurate representation of auditory cues along all f.

Similar to the previous cues, the IC model generates a spikes matrix $\mathbf{S}^{\mathbf{IC}}_{\mathbf{\Delta t}}$ at every time step Δt. Again, output angles can be computed with the same procedures applied to $\mathbf{S}^{\mathbf{MSO}}_{\mathbf{\Delta t}}$. Chapter 2.2 provides further details on the architecture of the MSO, LSO and IC models. Now we proceed to introduce and justify the selection of clustering methods.

4.2.3 Clustering of Spatial Cues

Clustering algorithms can be used directly for classification when having the same number of prototypes $p \in \{1, 2 \dots P\}$ and target classes $c \in \{1, 2 \dots C\}$. However, with a larger P it is possible to cover more closely the distribution underlying the training data, hence, improving the overall performance of the system. In the case of SSL, the distribution of auditory cues in each representation space can be highly convoluted. Therefore, using $P \gg C$ can spread the trained prototypes closer to the distribution of the characterised cue.

Since several p can belong to a single c, an additional requirement is the inclusion of another layer in the architecture for classifying the winning c. Again, the criteria for selecting clustering algorithms is to compare a standard statistical technique against a neural method, for which we choose K-Means (KM) (MacQueen, 1967), Learning Vector Quantisation (LVQ) (Lloyd, 1982; Kohonen, 1995) and Self Organising Feature Maps (SOM) (Kohonen, 1982, 2013).

4.2.3.1 K-Means Clustering

Due to its simplicity and speed relative to other clustering techniques, K-Means Clustering (KM) (MacQueen, 1967; Lee & Choi, 2010) is included as a benchmark against the more sophisticated SOM. The best results are achieved with $K = 26$ and using a randomly chosen sample of the training data as starting positions for the prototypes. $K = 26$ comes from the set of multiples of the total number of target classes C (sound source angles used during training), i.e., $K \in \{C \times 1, C \times 2, C \times 3, C \times 4, C \times 5\}$. All analytical procedures described in this chapter

Value	Variable	Comments
K	26	{13, 26, 39, 52, 65}
Distance	Squared Euclidean	{Squared Euclidean, City Block, Cosine, Correlation, Hamming}
Empty action	Singleton	{Error, Drop: Remove alone prototypes, Singleton: Create single instance cluster}
Online phase	On	{On: guarantees local minima, Off: slower}
Replicates	10	Times to repeat the clustering with new initial prototype positions. Take the solution with the lowest value for within-cluster sums of point-to-centroid distances.
Starting positions	Sample	{Sample: From data, Uniform: From data range, Cluster: Preliminary clustering with 10% of data}

Table 4.1: K-Means Clustering (KM)

use the Euclidean distance as the standard metric. For further details on the hyper-parameters see Table 4.1.

4.2.3.2 Learning Vector Quantisation

The Learning Vector Quantisation (LVQ) (Lloyd, 1982) classification method represents a step between K-Nearest Neighbours and Self-Organising Maps. In our experiments, we used the LVQ-2 variant, where the presented instance attracts the winning prototype and repels the second winner.

Value	Variable	Comments
Hidden units	26	{13, 26, 39, 52, 65}
Learning rate	0.01	{0.1, 0.01, 0.001, 0.0001}
Learn function	LV2	{LV1: Attracts winner, LV2: Attracts winner and repels second winner}
Train epochs	1000	{10, 100, 1000, 10000}
Train time	Inf.	Maximum training time.
Goal	0	Desired error.

Table 4.2: Learning Vector Quantisation (LVQ)

Value	Variable	Comments
X prototypes	13	{13, 26, 39, 52, 65}
Y prototypes	13	{13, 26, 39, 52, 65}
Map dimensions	2D	Map is projected on a 2D space.
Topology function	Hexagonal	{Square grid, Hexagonal, Triangular, Random}
Train epochs	1000	{10, 100, 1000, 10000}
Train time	Inf.	Maximum training time.

Table 4.3: Self Organising Map (SOM)

4.2.3.3 Self Organising Map

Due to its topology-preserving property, Self Organising Maps (SOM) (Kohonen, 1982) facilitate visualisation of the data structure in lower dimensions. We use a 2D SOM in two different configurations. In the first one $P = C$ and its output can be directly used for motor control. In the second configuration $P = C^2$ and a classification layer is added on top of it. In both cases the ordering phase consists of 1000 steps, has a learning rate $\eta = 0.9$ and the neighbourhood distance (ND) decreases from the furthest neurone to 1. The tuning phase consists of additional 4000 steps where $\eta = 0.02$ and $ND = 1$. For further details on the hyper-parameters see Table 4.3.

4.2.4 Classification of Spatial Cues

In our testing architecture, the classification layer receives input from the representation layer or an intermediate clustering layer. Following the same heuristic, we compare a standard statistical technique for benchmarking against a pair of artificial neural networks (ANN). K-Nearest Neighbours (KNN) (Cover & Hart, 1967) is the chosen statistical technique and the selected ANNs are the Radial Basis Functions network (RBF) (Park & Sandberg, 1991) and the Multilayer Perceptron (MLP) (Rosenblatt, 1958).

4.2.4.1 K-Nearest Neighbours

K-Nearest Neighbours (KNN) (Cover & Hart, 1967) is a relatively simple, yet powerful, classification technique. Instead of exhaustive search, we use a KD-Tree to reduce the cost of finding the nearest neighbour from $O(N^2)$, to $O(NlogN)$ for N data points (Bentley, 1975). We obtained the best performance with $K = 4$.

4.2.4.2 Radial Basis Functions Network

A significant advantage of Radial Basis Functions Networks (RBF) (Park & Sandberg, 1991) over other ANNs is their much faster training procedure. The number of neurones in the hidden layer is equal to the number of training instances, and the network shows the best overall performance with a spread $\sigma = 10$.

4.2.4.3 Multilayer Perceptron

The Multilayer Perceptron (MLP) (Rosenblatt, 1958) is a universal function-approximator robust to noise, whose internal dynamics are one of the best understood in the field of ANNs. During training we use the following data ratios: *training* $= 0.8$, *validation* $= 0.1$ and *testing* $= 0.1$. We use hyperbolic tangent as activation function and, due to its increased speed for large networks, we use the scaled conjugate gradient (Møller, 1993) method for backpropagation. The optimisation parameters are set to the standard values $\sigma = 5 \times 10^{-5}$ and $\lambda = 5 \times 10^{-7}$ according to Møller (1993).

Value	Variable	Comments
Hidden units	26	{13, 26, 39, 52, 65}
Train ratio	0.8	Proportion of data for training.
Valid ratio	0.1	Proportion of data for crossvalidation.
Test ratio	0.1	Proportion of data for testing.
Train function	SCG	Scaled conjugate gradient.
Train epochs	1000	
Train time	Inf.	
Goal	0	Desired error.
Min gradient	1e-60	Desired gradient in error landscape.
Max fails	500	Maximum epochs with lower performance.
Sigma	5e-05	Determines the change in the weight for the second derivative approximation. Optimisation value recommended by Møller (1993).
Lambda	5e-07	Regulates the indefiniteness of the Hessian. Optimisation value recommended by Møller (1993).

Table 4.4: Multilayer Perceptron (MLP)

The number of hidden neurones (HN) changes depending on the architecture being tested. When the MLP receives input from the representation layer $HN = \lfloor \mathbf{v}_{\text{dim}}/2 \rceil$, where $\mathbf{v}_{\text{dim}}$ is the dimensionality of input vector $\mathbf{v}$. When the MLP receives input from the clustering layer $HN = C \times 2$. For further details on the hyper-parameters see Table 4.4.

4.2.5 System Performance

The system performance is analysed using measures from information retrieval theory (Van Rijsbergen, 1979): *recall* (Re), *precision* (Pr), *specificity* (Sp), *accuracy* (Ac) and *F-measure* (Fm). We compute the value of each measure from the confusion matrices of the output angles. Specifically from the *true positives*

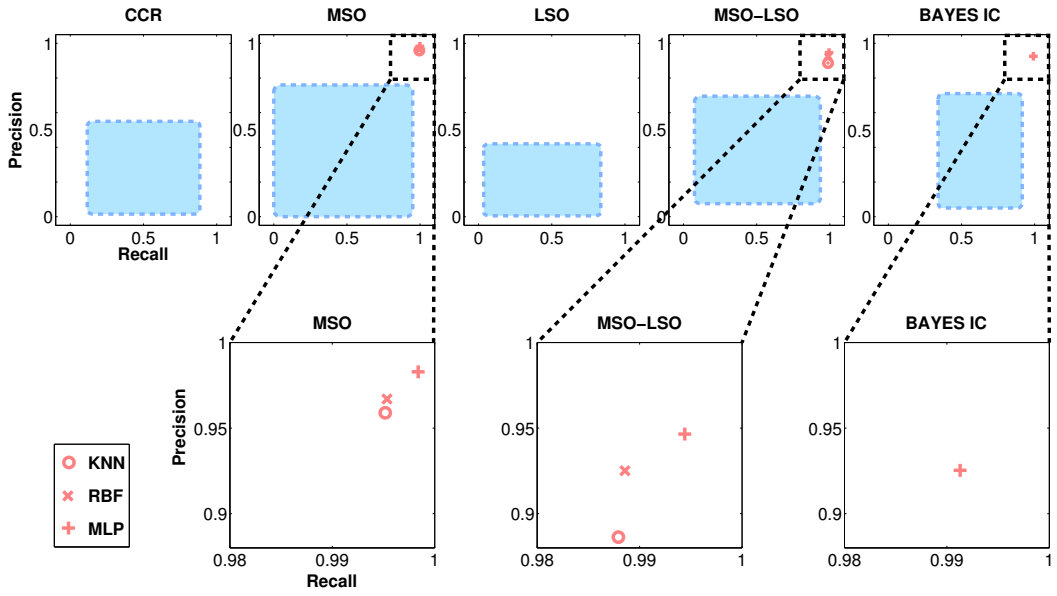


Figure 4.3: Recall - Precision. The markers show the performance of the best representation methods. For clarity, the area where the performance of the non-winning classifiers falls is represented by shaded squares in the top five plots. In all representations the best performance is achieved by the *training / testing* configuration *Speech / WN*. Three representation methods showed significantly better performance: MSO, MSO-LSO and IC. Within these representations, three classification algorithms obtain best results with a recall $Re > 0.98$ and precision $Pr \geq 0.89$: KNN, MLP and RBF.

(TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN):

$$Pr = \frac{TP}{TP + FP} \ , \tag{4.5a}$$

$$Re = \frac{TP}{TP + FN} \ , \tag{4.5b}$$

$$Sp = \frac{TN}{TN + FP} \ , \tag{4.5c}$$

$$Ac = \frac{TP + TN}{TP + TN + FP + FN} \ , \tag{4.5d}$$

$$Fm = 2 \times \frac{Pr \times Re}{Pr + Re} \ . \tag{4.5e}$$

Table 4.5: K-NN - Classification performance with each representation method. Best results are highlighted in bold.

	Pr	Re	Sp	Ac	Fm
CCR	0.21	0.75	0.59	0.61	0.33
MSO	**0.96**	**1.00**	**0.96**	**0.98**	**0.98**
LSO	0.15	0.65	0.54	0.55	0.25
MSO-LSO	**0.89**	**0.99**	**0.90**	**0.94**	**0.93**
Bayes IC	0.18	0.75	0.53	0.56	0.29

Table 4.6: MLP - Classification performance with each representation method. Best results are highlighted in bold.

	Pr	Re	Sp	Ac	Fm
CCR	0.55	0.89	0.75	0.79	0.68
MSO	**0.98**	**1.00**	**0.98**	**0.99**	**0.99**
LSO	0.21	0.76	0.56	0.59	0.33
MSO-LSO	**0.95**	**0.99**	**0.95**	**0.97**	**0.97**
Bayes IC	**0.93**	**0.99**	**0.94**	**0.96**	**0.96**

4.3 Experimental Results

Appendix A contains detailed results for each variant defined by the testing architecture. The performance of all representation and classification algorithms is displayed in recall-precision plots in Figure 4.3. In all representations the top performance is achieved when *training / testing* with *Speech / WN*. From all the tested representation methods, three lead to much more accurate results: MSO, MSO-LSO and Bayes IC. Furthermore, within those three representations, three classification algorithms perform significantly better than the rest with $Re > 0.98$ and $Pr \geq 0.89$: KNN, MLP and RBF.

The performance measures of the three best, or winning, classifiers is shown in Tables 4.5, 4.6 and 4.7. In order to show the considerable increase in performance of the winning classifiers, Table 4.8 shows the performance results of the *second best* classifiers with $Pr > 0.7$. These *second best* systems achieved higher

Table 4.7: RBF - Classification performance with each representation method. Best results are highlighted in bold.

	Pr	Re	Sp	Ac	Fm
CCR	0.20	0.75	0.56	0.58	0.32
MSO	**0.97**	**1.00**	**0.97**	**0.98**	**0.98**
LSO	0.42	0.83	0.70	0.73	0.56
MSO-LSO	**0.93**	**0.99**	**0.93**	**0.96**	**0.96**
Bayes IC	0.27	0.73	0.68	0.68	0.40

Table 4.8: Performance of the *second best* systems with $Pr > 0.7$. The same classifier has better performance when clustering input from the MSO.

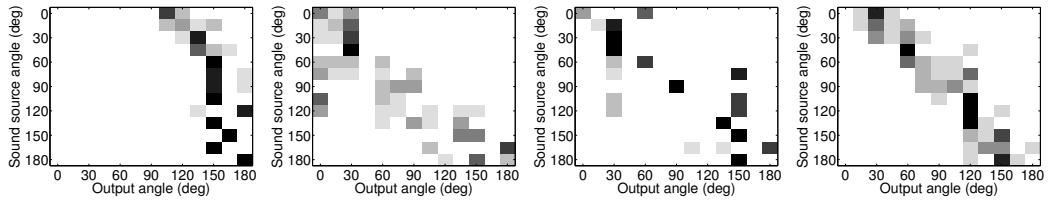
	Pr	Re	Sp	Ac	Fm
MSO: KM-RBF	0.75	0.94	0.85	0.88	0.84
MSO: SOM-RBF	0.75	0.94	0.85	0.88	0.83

performance when clustering and classifying input from the MSO representation.

In the following subsections we detail the performance of the best classifiers with each of the cue representations, and compare them against the WTA classification rule we applied in our previous work with a different robotic platform (Dávila-Chacón *et al.*, 2012). In all cases the *training / testing* configuration is *Speech / WN*.

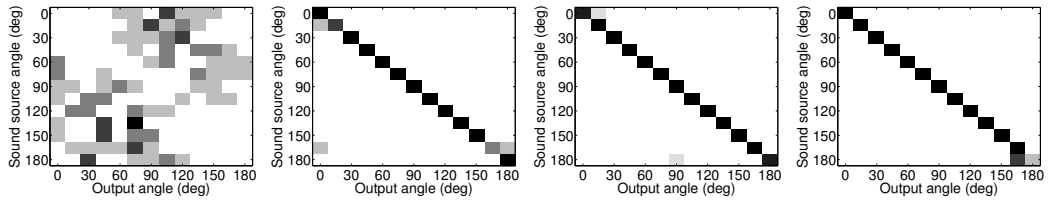
4.3.1 Cross Correlation

This statistical method shows lower performance than the MSO model, its bioinspired counterpart. However, the confusion matrices in Figure 4.4 show that the angle deviation from the ground truth of KNN and RBF outputs is small for practical purposes when using CCR as input. It remains an open possibility to improve the performance of this method when adding a noise cancelling layer to the system. This enhancement is desirable for online applications as CCR provides a faster characterisation of ITDs than the MSO.



(a) CCR - WTA **(b)** CCR - KNN **(c)** CCR - MLP **(d)** CCR - RBF

Figure 4.4: Confusion matrices when using CCR representation as input for WTA and the winning classification methods.



(a) MSO - WTA **(b)** MSO - KNN **(c)** MSO - MLP **(d)** MSO - RBF

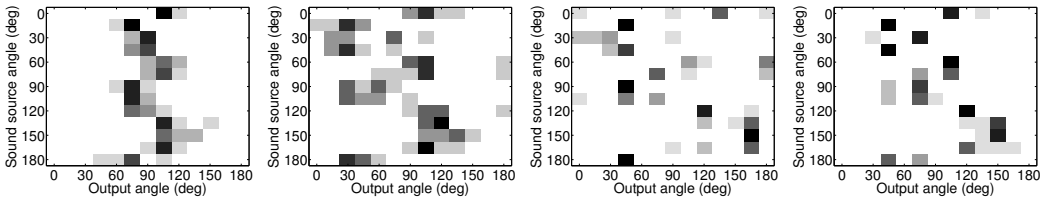
Figure 4.5: Confusion matrices when using MSO representation as input for WTA and the winning classification methods.

4.3.2 Medial Superior Olive Model

The MSO allows the system to reach the highest accuracy relative to all other representations. Also, it is the only representation that allows the three winning classification methods to perform almost flawlessly. Figure 4.5 shows the improvement of the winning classifiers over the baseline method of WTA. The MSO performed robustly under high levels of ego-noise, even when the noise frequency components were overlapping with the f provided by the PHFB.

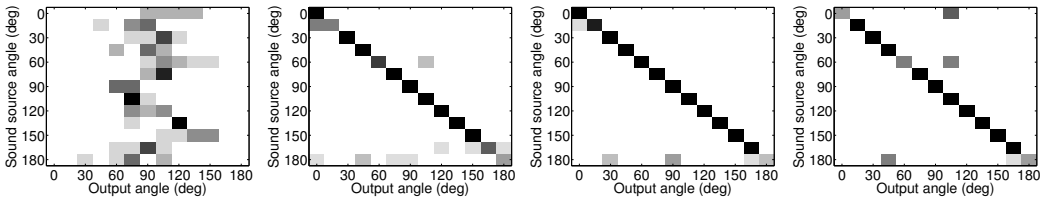
4.3.3 Lateral Superior Olive Model

This bioinspired method is the only one we used for representing ILDs, as there are no standard statistical techniques for benchmarking. The extraction of ILDs is affected by the geometrical and material properties of the robotic head. In previous work the authors successfully used ILDs for SSL (Liu *et al.*, 2010) with



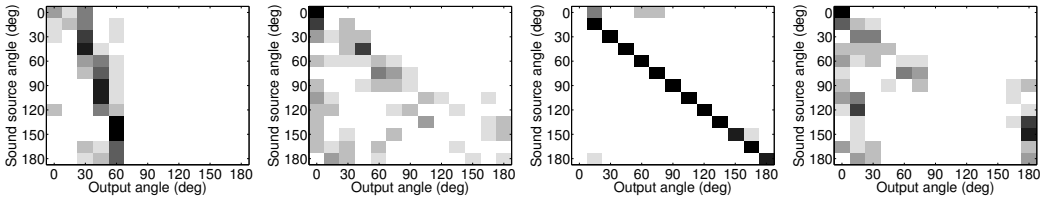
(a) LSO - WTA **(b)** LSO - KNN **(c)** LSO - MLP **(d)** LSO - RBF

Figure 4.6: Confusion matrices when using LSO representation as input for WTA and the winning classification methods.



(a) MSO - LSO - WTA **(b)** MSO - LSO - KNN **(c)** MSO - LSO - MLP **(d)** MSO - LSO - RBF

Figure 4.7: Confusion matrices when using MSO and LSO representations as input for WTA and the winning classification methods.



(a) Bayes IC - WTA **(b)** Bayes IC - KNN **(c)** Bayes IC - MLP **(d)** Bayes IC - RBF

Figure 4.8: Confusion matrices when using Bayes IC representation as input for WTA and the winning classification methods.

a styrofoam humanoid head. Nevertheless, Figure 4.6 shows that the classification techniques can not infer correctly the location of sound sources from ILDs extracted with the iCub head.

One possible reason for this low performance is the presence of high levels of ego-noise. After inspecting a spectrogram of the iCub's ego-noise, we found

that the spectral region with the most intense ego-noise contains15 of the 20 frequency components f defined by the PHFB cochlear model. Therefore, noise in these frequencies can significantly reduce the SNR of incoming stimuli and impede the use of ILDs for SSL. Subsection 4.2.1 provides more details on the PHFB preprocessing step.

Another possibility is that the inaccuracy of the system when using ILDs is due to the material properties of the robotic head. A difference from the platform used by (Liu *et al.*, 2010) is that the iCub head is hollow and has openings in the back, reducing in this way the shadowing effect needed to use ILDs for SSL effectively.

4.3.4 Linear Integration of Time and Level Differences

This integration of ITDs and ILDs, represented by the MSO and LSO models, is much simpler than the IC Bayesian integration. In this case, the MSO and LSO activation matrices for Δt are merely appended and used as input for the next system layer. It is interesting to see in Figure 4.7 that the performance of the three winning algorithms dramatically increase in comparison to the IC method, even though the complexity of the characterisation procedure decreases. However, it is also important to keep in mind that the training procedure also becomes more demanding as the dimensionality of the input vectors to the classification layer grows by a factor of $\sim$7.

4.3.5 Bayesian Integration of Time and Level Differences

This representation is the most biologically plausible from the set we describe in this chapter, but it is also the most computationally expensive. On the other hand, the dimensionality reduction provided by this method speeds up considerably the training procedure of the classification algorithms.

Figure 4.8 shows the confusion matrices when using WTA for classification, versus the performance of the three winning classifiers. The output of WTA is strongly biased towards a small range of angles on the 0° quadrant, possibly due to the non-linear encoding of information across the IC neurones. Also, KNN

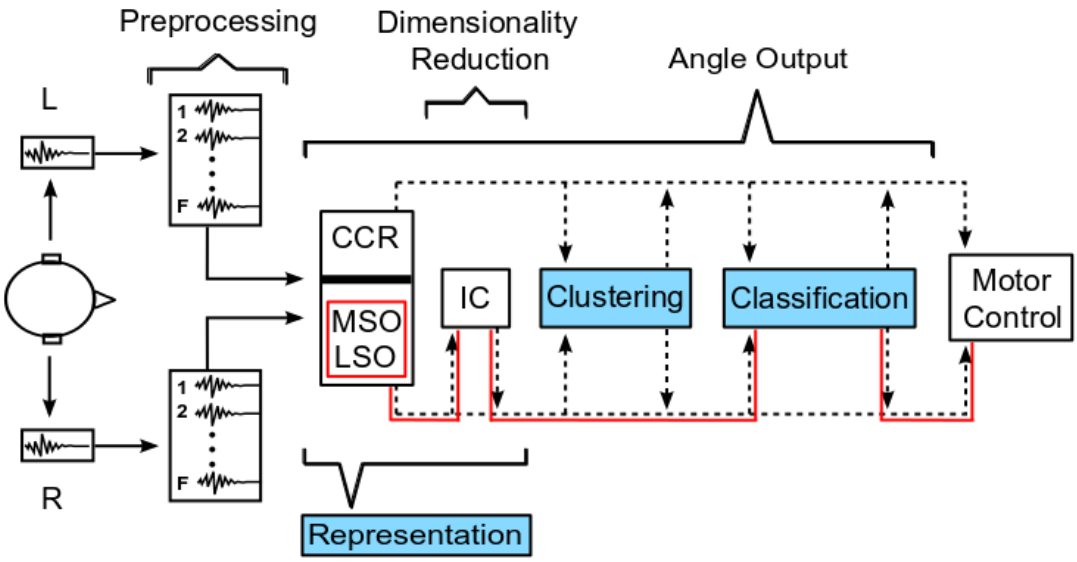


Figure 4.9: Sound source localisation testing architecture. The red path indicates the best performing system. Sound pre-processing consists in decomposing the sound input in several frequency components with the Gammatone filterbank emulating the human cochlea Slaney (1993). Afterwards, the MSO and LSO models *represent* ITDs and ILDs respectively. The IC model integrates output from the MSO and LSO while performing *dimensionality-reduction*. Finally, the *classification* layer produces an output angle that is used for motor control.

and RBF show a bias, albeit smaller, towards the same region. In contrast, the MLP is capable of correctly encoding the spiking activity of the IC.

4.4 Conclusion

After our extensive comparison of architectures, we found the winning variant that Figure 4.9 indicates with the red path. The increase in performance is better understood when we analyse the advantages of the components that performed better at each stage. In principle, it is possible to extract ITDs from any pair of microphones separated by a known distance. Such a configuration can work with or without a humanoid head between the microphones. However, for the estimation of ILDs, it is necessary to measure the shadowing effect produced by a head-like structure. The results of the static SSL experiments show that the

iCub's head produces a similar shadowing to the Soundman's head, independently of their structural and material differences. Soundman is specifically designed for the generation of spatial effects in binaural recordings; hence, these results support the use of iCub for experiments in spatial audition. In this chapter, we compare different methods for the representation and classification of spatial cues for SSL. We found three best representation methods: MSO, MSO-LSO and Bayes IC. There are also three winners from our set of classifiers: KNN, MLP and RBF.

The fastest method for representation of ITDs is the MSO model alone. Nevertheless, MSO-LSO and Bayes IC methods can be more robust when classifying sounds richer in high-frequency components. We have shown that the LSO model performs well under lower levels of ego-noise (Dávila-Chacón *et al.*, 2012). More precisely, with levels of ∼40 Hz instead of ∼60 Hz. An interesting direction for future work is to test the system using ILDs in combination with a noise cancelling module, as we expect this configuration will improve the accuracy of SSL with the iCub head. Concerning training speed, the fastest classification method is KNN. However, for life-long learning, the standard KNN method would become computationally expensive, i.e. slow, as the system would need to store a vast number of prototypes from possibly several environments. Therefore, the MLP and RBF networks represent a more practical option regarding online speed.

Finally, an exciting extension of the system is to include the propagation of probabilities through time and to increase the confidence of the sound source angle by integrating vision (Natale *et al.*, 2002; Lv & Zhang, 2008). We expect that both additions will improve the confidence of the classification algorithms and their robustness against higher levels of reverberation.

Chapter 5

Dynamic Automatic Speech Recognition

In the previous chapters, we have dived into sound source localisation, starting with the biological principles and ending with the design of a biomimetic system for humanoid robots. At this point our focus shifts to the application of our increased understanding of SSL towards the improvement of automatic speech recognition under the same noisy conditions. In Section 5.2 we detail an experiment where we test the performance of ASR with two humanoid robots. We assume that the combination of SSL and ASR can lead to increased accuracy when the robot orients its face at an optimal angle from the sound source. In Section 5.3 we present an experiment whose objective is to find how many SSL iterations it takes the system to face a sound source when the robot starts from a range of different angles between the sound source and the direction faced by the robot. Once the robot is facing directly at the sound source, we can measure the stability of the SSL system for locking on the speech target. We refer to this scenario as dynamic SSL, and it is an essential test in real-world situations, where potential outliers in SSL can disrupt the effect of engaged communication with a human subject.

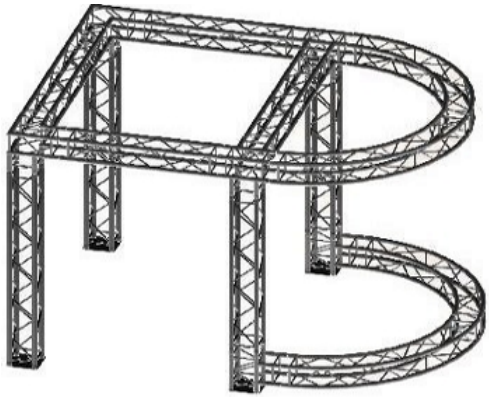


Figure 5.1: Scaffold of the audio-visual VR experimental setup surrounding the subject. Reprinted with permission from Bauer *et al.* (2012) (Copyright © 2012, IEEE).

5.1 Smoke and Mirrors

In the following subsections, we introduce the robotic platforms that we use. We compare the performance of ASR with the robot iCub and a dummy head designed to perform binaural recordings that maximise the spatial effect for human listeners. Taking advantage of the versatility of the virtual reality setup that we designed for our previous experiments in SSL, we place both platforms inside a semi-circle of loudspeakers from where we reproduce the auditory stimuli (Klein *et al.*, 2000). Our experiments are designed to measure the effect that the materials and geometry of both platforms may have on the performance of ASR; hence, we repeat the experiment with each platform.

5.1.1 Virtual Reality Setup

The virtual reality (VR) setup that our group designed is equipped to test multimodal-integration architectures and to present visual and auditory stimuli to robotic and human subjects. The VR setup allows the user to control the spatial and temporal production of images and sounds in a semi-circular projection screen. Figure 5.1 shows the structure surrounding the subject and Figure 5.2 the setup of the image projectors. In our first experiments, we focus on what we call *static ASR*, as both the speech source and the robot orientation remain

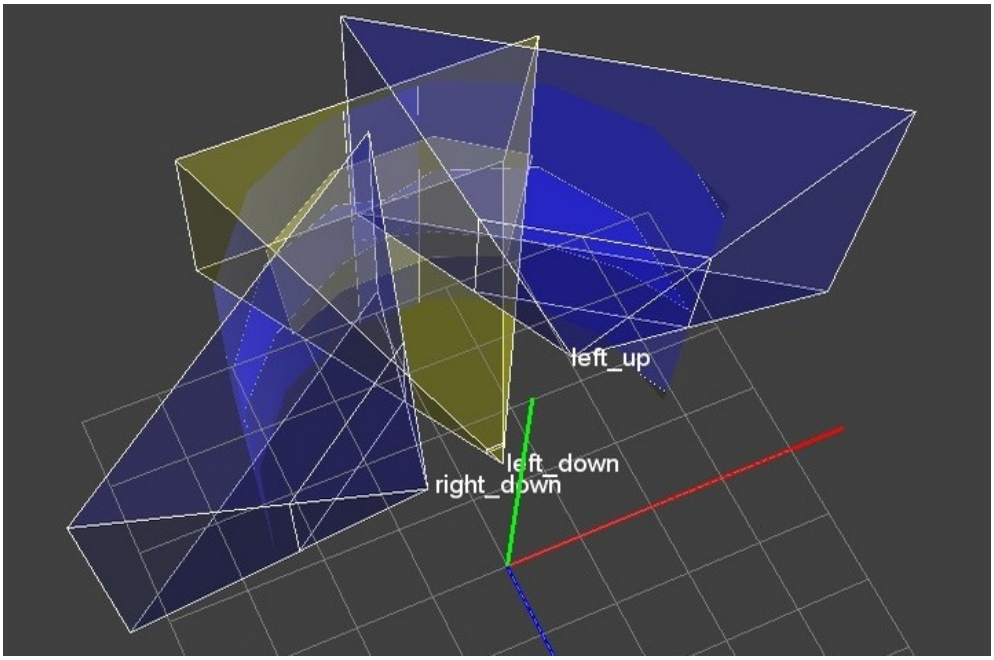


Figure 5.2: Setup of the projectors in the audio-visual VR experimental setup. Reprinted with permission from Bauer *et al.* (2012) (Copyright © 2012, IEEE).

fixed during the entire experiment. This constraint intends to avoid the interference of additional sources of ego-noise that can affect the performance of ASR while performing SSL (Perisa *et al.*, 2004; Barker *et al.*, 2005).

The primary objective is to measure the effect that the incidence angle of sound waves has on ASR. As the face and pinnae of the robot obstruct and reflect the sound waves differently, depending on the incidence angle of the sound stimuli, this may enhance or reduce the SNR of speech signals. When we run the experiments, we place the humanoid at the radial centre of a projection screen shaped as a half cylinder. As we see in Figure 5.3, the humanoid is located at the radial centre of a projection screen shaped as a half cylinder and the noise produced by the projectors is below 30 dB at the location of the robot. This setup allows us to test the system performance down to a granularity of 0°, as behind the screen there are 13 speakers evenly distributed on the azimuth plane at angles $\theta_{lspk} \in \{0°, 15°, \ldots, 180°\}$. The loudspeakers lay on a circumference with a radius of ∼1.6 m around the robot. Corrugated curtains partially damp the room acoustics in order to approach a reverberation time ($0.25 - 0.5$ s) and an

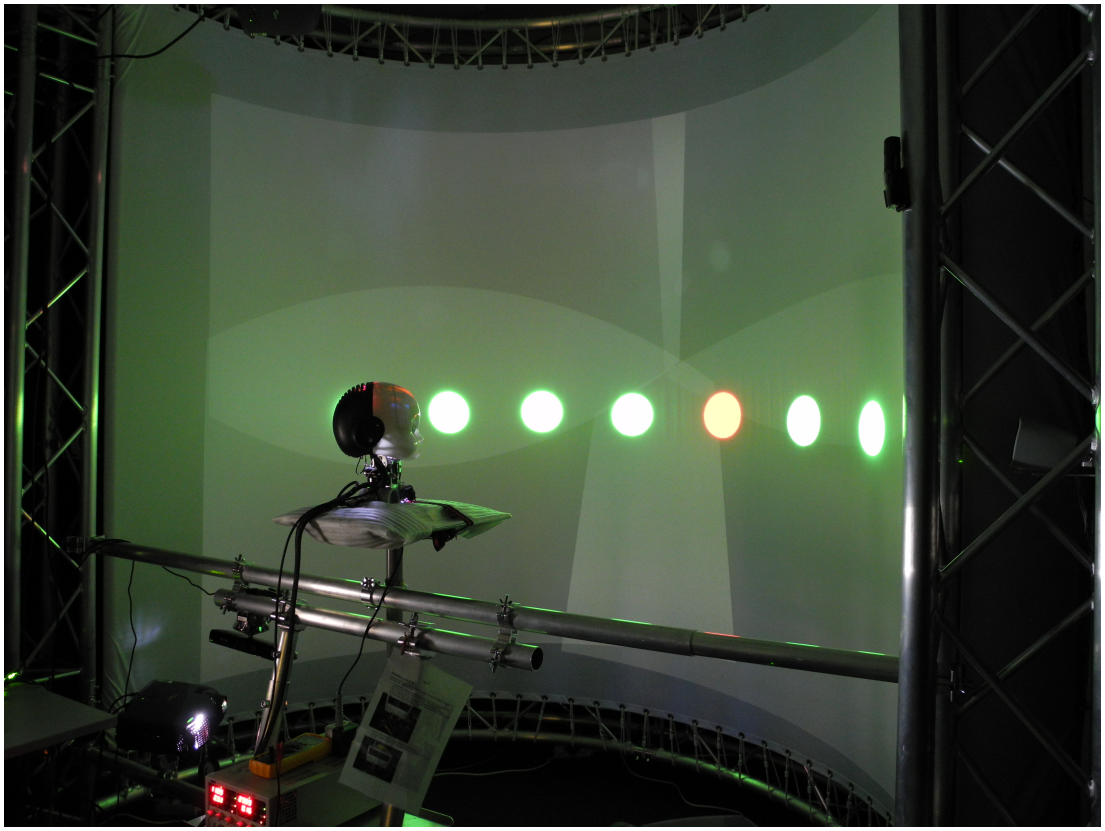


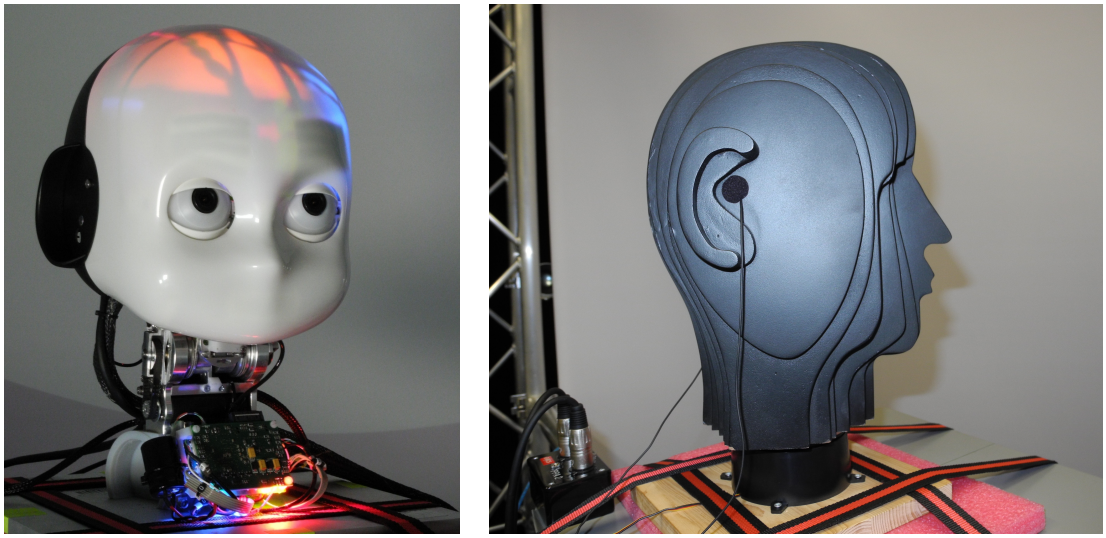
Figure 5.3: Setup of the loudspeakers in the audio-visual VR experimental setup. The dots represent the location of sound sources behind the screen.

inner sound pressure level ($20-40$ dB) with *studio* quality. We provide a detailed description of this setup and the principles behind its design can in Bauer *et al.* (2012).

5.1.2 Humanoid Robotic Platforms

The humanoid platforms used in our experiments are the iCub robotic head (Beira *et al.*, 2006) and the Soundman wooden head[1] modified by our group to rotate on the azimuth plane with the help of an electric motor installed in its base. The iCub is a humanoid robot designed for research in embodied embedded cognition (Metta *et al.*, 2008) and in cognitive developmental robotics (Asada *et al.*,

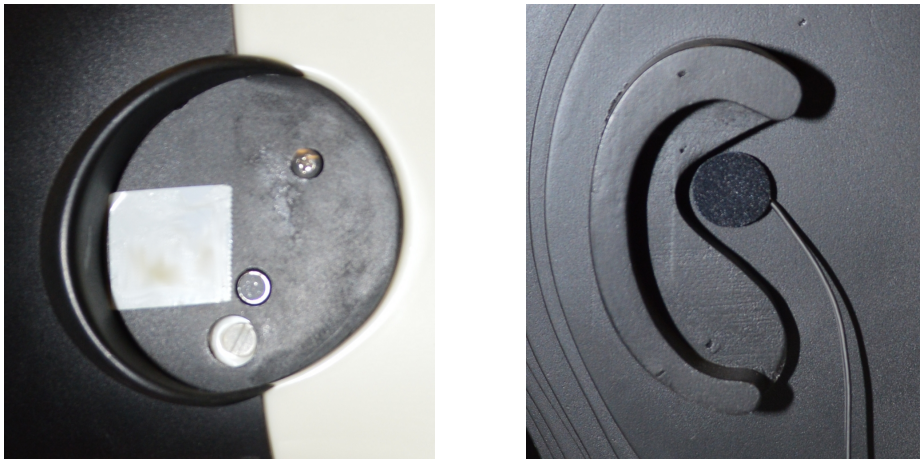
[1] `http://www.soundman.de/en/dummy-head/`



(a) iCub **(b)** Soundman

Figure 5.4: Humanoid heads used in our experiments.

2001). Soundman is a commercial device, with the geometry and dimensions of a human head, designed for the production of binaural recordings that maximise the perception of spatial effects. Both platforms offer the possibility of extracting spatial cues from binaural sound, as the geometric and material properties of both humanoid heads (Hwang *et al.*, 2006) produce interaural time and level differences. Our goal is to find out if the resonance of the iCub head, from the skull and interior components, has an impact on the performance of ASR. When we perform ASR experiments with the iCub *Off* or when we use Soundman, we mount the same pair of balanced microphones on either head and control the sound stimuli to have an intensity of ∼60 dB. When we perform SSL experiments with the iCub *On*, we increase the intensity of the sound stimuli to ∼80 dB due to the high levels of ego-noise produced by the robot. Both platforms can be seen in Figure 5.4 and their respective pinnae is shown in Figure 5.5. Notice that the pinnae differ considerably between each other. Nevertheless, the functional aspect of such artificial shapes is the creation of asymmetries that can create a unique imprint by absorbing in different magnitudes the frequency components of incoming sound waves (Hofman *et al.*, 1998; Finger *et al.*, 2010).



(a) iCub **(b)** Soundman

Figure 5.5: The robots ears consist of microphones perpendicular to the sagittal plane, surrounded by *asymmetrical* pinnae made of plastic (iCub) or *symmetrical* pinnae made of wood (Soundman).

5.2 Robot Speech Recognition

In this first experiment of the chapter, we test this hypothesis with experiments to find the orientation that increases the accuracy of ASR and measure the magnitude of the improvement in comparison to other angles, if any. More formally, we present an embodied embedded cognition approach to ASR, where the humanoid robot adjusts its orientation to the angle that increases the signal-to-noise ratio of speech. In other words, the robot turns its face to 'hear' better the speaker in a similar way as some elders or people with auditory deficiencies do.

5.2.1 Speech Recognition and Phonetic Post-Processing

We use a system developed by our group for automatic speech recognition Twiefel *et al.* (2014): Domain- and Cloud-based Knowledge for Speech Recognition (DOCKS). The DOCKS system has two main components: 1) A *domain-independent* speech recognition module and 2) a *domain-dependent* phonetic post-processing module. The need for domain-dependent ASR arises from the intense noise of the cooling system in humanoid platforms commonly used for research in academia (Nao,

Table 5.1: Performance of ASR systems.

System	WER in %	SER in %
Go	50.230	97.804
Sp + NG	60.462	95.101
Sp + FSG	65.346	85.980
Sp + DoSe	65.346	85.980
Go + Sp-HMM + NG	7.962	27.703
Go + Sp-HMM + FSG	6.038	19.257
Go + Sp-HMM + DoSe	5.846	18.581
Go + WoLi	23.231	57.432
Go + SeLi (**DOCKS**)	**3.077**	**11.993**

Best results are marked in boldface.

Terminology can be found in the text.

iCub). In such conditions, sentences are more easily recognisable than words, which is analogous to the RAF alphabet used in aviation to communicate under low SNR conditions. The domain-dependent output of the DOCKS system does not impede generalisation from our experimental results, as our objective is not to develop a novel ASR system. Our goal is to compare the performance of any existing ASR system *with* and *without* the support of SSL.

To test the DOCKS ASR system Heinrich and Wermter created a corpus that contains 592 utterances produced from a predefined grammar Heinrich & Wermter (2011b). The corpus was recorded by female and male non-native speakers using headset microphones, and it is especially useful as the grammar for parsing the utterances is available. They use two commercial ASR platforms as the domain-independent component of the DOCKS system: Google ASR Schalkwyk *et al.* (2010) and Sphinx Walker *et al.* (2004). Afterwards, they measure the word (WER) and sentence (SER) error rates under four different configurations. In Table 5.1 we compare the performance of 1) the raw output of *Google ASR* (Go), 2) *Sphinx ASR* (Sp) with an N-Gram language model (NG), with the corpus

finite state grammar (FSG) and with the domain sentences (DoSe), 3) Go plus the *Sphinx Hidden Markov Model* (Sp-HMM) with NG, with FSG and with DoSe, and 4) Go with the domain word list (WoLi) and with the domain sentence list (SeLi).

During the *domain-independent* speech recognition, the DOCKS system uses *Go* as in previous work Rubruck *et al.* (2013a) it has shown better performance than *Sp*. In our experiments, we use the TIMIT core-test-set (TIMIT-CTS) Garofolo *et al.* (1993) as speech stimuli. The TIMIT-CTS is formed by the smallest TIMIT subset that contains all existing phonemes in the English language. It consists of 192 sentences spoken by 24 different speakers: 16 male and 8 female pronouncing 8 sentences each. Further details about the DOCKS architecture can be found in Twiefel *et al.* (2014) and Davila-Chacon *et al.* (2013).

During the *domain-dependent* phonetic post-processing, the DOCKS system maps the output of *Go* to the sentences in the TIMIT-CTS. When the system sends a sound file to *Go*, it returns the 10 most plausible sentences (G10). First, the system transforms the G10 and the TIMIT-CTS from *grapheme* representation to *phoneme* representation Bisani & Ney (2008). Then the system computes the Levenshtein distance Levenshtein (1966) between each of the phoneme sequences in the G10 and the TIMIT-CTS. Finally, the phoneme sequence in the TIMIT-CTS with the smallest distance to any of the phoneme sequences in the G10 is considered the winning result. We consider correct the sentence corresponding to the winning phoneme sequence when it matches the ground truth sentence presented to the robot.

In general, the Levenshtein distance $\mathcal{L}(\mathbf{a}, \mathbf{b})$ refers to the minimum number of *deletions*, *insertions* and *substitutions* required to convert string $\mathbf{a} = a_1, \ldots, a_i, \ldots, a_m$ into string $\mathbf{b} = b_1, \ldots, b_j, \ldots, b_n$. We compute the distance $\mathcal{L}(\mathbf{a}, \mathbf{b}) = \mathcal{D}(m, n)$ as follows:

$$\mathbf{D}(i,j) = \begin{cases} i & \text{for } \; 0 \leq i \leq m \; \text{ and } \; j = 0, \\ j & \text{for } \; 0 \leq j \leq n \; \text{ and } \; i = 0, \\ \min \begin{cases} \mathbf{D}(i-1, j) + 1 \\ \mathbf{D}(i, j-1) + 1 \\ \mathbf{D}(i-1, j-1) + \kappa \end{cases} & \text{for } \; 1 \leq i \leq m \; \text{ and } \; 1 \leq j \leq n. \end{cases}$$

Where $\kappa = 0$ if $a_i = b_j$ and $\kappa = 1$ if $a_i \neq b_j$.

5.2.2 Experimental Results

In the experiments, we present the speech stimuli around the robotic heads from the loudspeakers at angles θ_{lspk} at $\sim$1.6 m from the robot and measure the accuracy of the DASR system. Let θ_{neck} be the angle faced by the robot at any given time, and δ_{diff} be the angular difference between θ_{lspk} and θ_{neck}. We hypothesise that there is an angle -or set of angles- δ_{best} for which the signal-to-noise ratio (SNR) is highest and hence, for which the DASR system performs better. For this purpose, we measure the performance of the DASR system after reproducing 10 times the CTS utterances from the loudspeaker at angle θ_{lspk}. The performance is measured as the average success rate at the sentence-level for the entire CTS corpus over the 10 trials. We refer to success rate as to what *sentence accuracy* is in the ASR domain, i.e., the ratio of correct recognitions over the total number of trials. It is also desirable to visualise the results of this binary evaluation in a continuous domain (Liu & Shen, 2010b). We compute such transformation by measuring the average Levenshtein distance between the output of the DASR system and the ground truth sentences.

It is important to remember that the sounds recorded through the robotic heads contain 2 channels, i.e. the audio waves from the left and right microphones. As the DASR system requires monaural files as input, there are 3 possible reduction procedures: Using the sound wave from the left channel only (LCh), using the sound wave from the right channel only (RCh) or averaging the sound waves from both channels (LRCh). The average success rates of the 3 reduction procedures on the recordings obtained with both heads are shown in figure 5.6 and the average Levenshtein distances in figure 5.7. It is clear that the performance curves obtained from the recordings of both robotic heads follow the same patterns. Notice that the performance of the DASR system improves with the Soundman head on the most favourable angles δ_{best}. However, the difference is not significant enough to conclude that the resonance of the iCub head reduces the performance of the DASR system.

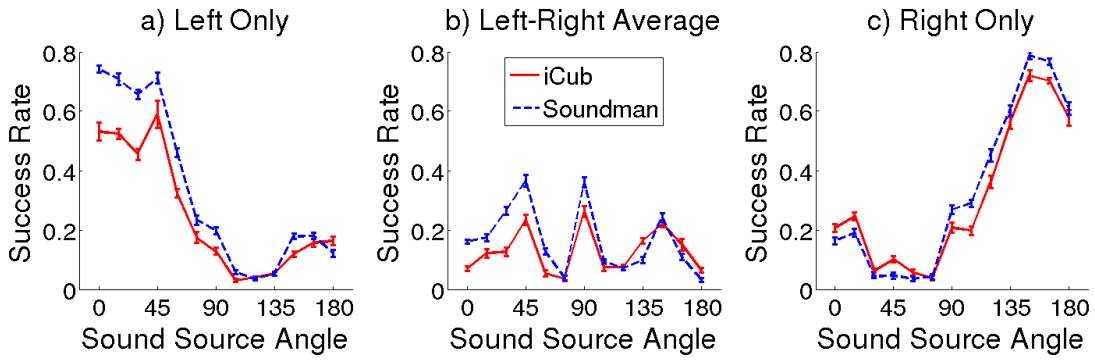


Figure 5.6: Average success rates of the DASR system. Results obtained with both robotic heads for the frontal 180° on the azimuth plane.

Even though we control the volume of each loudspeaker to output the same intensity level (∼75 dB), the smoothness of the performance curves is affected by the difference in fidelity from each of the loudspeakers. Nevertheless, the graphs clearly show a set of angles δ_{best} where the DASR system considerably improves its performance. For all reduction procedures with both robotic heads performance is best near $\delta_{best} \in \{45°, 150°\}$, where the robotic heads reduce the SNR of incoming speech minimally.

In the LRCh reduction, most sound source angles θ_{lspk} produce recordings where one channel has higher SNR than the other. Therefore, when we average both signals the speech SNR diminishes. The exceptions are sound sources at 90°, 45° and 150°. We conjecture that the moderate SNR that both channels have in the case of 90°, and the high and low SNR in the ipsilateral and contralateral signals in the case of 45° and 150° explain this discrepancy. It is also important to notice the magnitude of this effect, as the highest success rates from the LCh and RCh reductions are two times better than the highest success rates from the LRCh reduction. This difference can be related to the strong shadowing from the geometry and material of the humanoid heads. The same effect appears in the LCh and RCh reductions alone. We expected that the SNR of speech increases when the sound source is in front of the robot or parallel to the interaural axis, and when the input to the ASR system comes only from the channel closest to the sound source. However, the angles found to be optimal for our DASR system

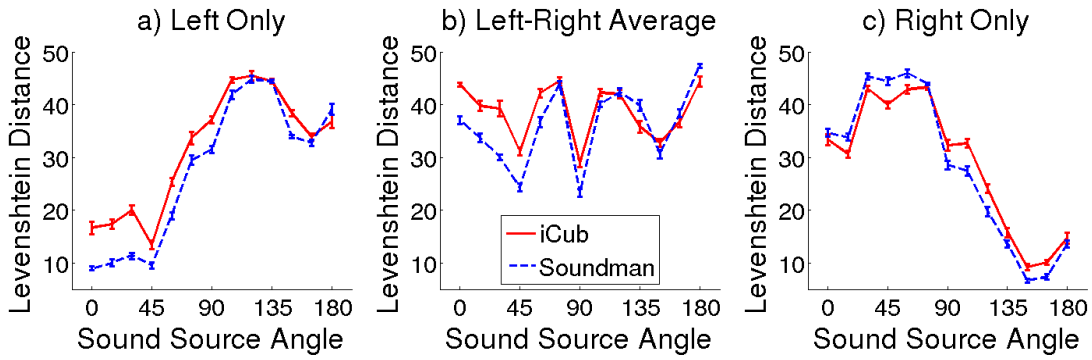


Figure 5.7: Average Levenshtein distances between the DASR output and the ground truth. Results obtained with both robotic heads for the frontal 180° on the azimuth plane.

are counter-intuitive, and the difference between the lowest and highest values in the LCh and RCh reductions is unexpectedly large.

The effect of the round shape of the heads and the position of the microphones explain the periodical shape in the LCh and RCh plots. The pinnae are placed slightly behind the coronal plane. Therefore, the distance travelled by the sound waves from the sound source to the contralateral ear is maximal at approximately 45° and 135° instead of 0° and 180°. This asymmetry explains the increase in performance before 135° for LCh and after 45° for RCh. On the other hand, the shadowing of the pinnae and reverberation from the metal structure on the sides of the VR setup could produce the decrease in performance before 45° for LCh and after 135° for RCh. Once the results of the first experiment are ready, we proceed to analyse the convergence of our system to the sound source after a sequence of localisation steps.

5.3 Acquisition Time and Source Locking

In this second experiment of the chapter, we test the hypothesis that our SSL system converges to the correct sound source angle in a short sequence of localisation steps. It is well understood the benefit of a humanoid appearance to enhance the human-robot interaction (HRI) (Mori, 1970; Minato *et al.*, 2004) and particularly the advantages of SSL for HRI (Lee *et al.*, 2009). but there are additional

metrics that have been proposed to asses the effectiveness of HRI (Goodrich & Schultz, 2007). A particularly relevant metric to asses the quality of engagement in a dialogue, is the effect of capturing the attention of both parties (*acquisition time*) and holding the interest of the human and the robot (*duration*) (Steinfeld *et al.*, 2006). When we say that SSL can help to improve the performance of automatic speech recognition, we assume that the robot will turn to the optimal listening angle in a small number of *SSL iterations*, what reflects the HRI metric of acquisition time. Once the robot is optimally oriented, it should remain stable in such position or proceed to track the speech source closely as soon as the source moves around the robot. We refer to this behaviour as *locking*, what reflects the HRI metric of duration.

5.3.1 Compound Stimuli and Convergence to Source

During the experiment, we measure the SSL locking on each of the 13 loudspeakers in the VR setup, at angles θ_{lspk}, in order to verify that the SSL system is robust to the reverberation produced in different room locations around the robot. As stimuli, we present the robot with a sound composed of utterances from 24 different speakers: 16 males and 8 females. More specifically, we append the longest sentence from each speaker in the TIMIT-CTS corpus in a single sequence of utterances to form a *compound sound* with a duration of 106 seconds. Once we form a compound sound, we move the last two sentences of the sequence of utterances to the beginning, creating in this way a set of compound sounds. By repeating this procedure, we create a total of 12 compound sounds. The objective of this method is to discard the possibility that the voice of a particular speaker systematically affects the SSL system at the same point in time.

At the beginning of each trial, the robot turns to a starting neck angle $\theta_{neck} \in \{45°, 15°, \ldots, 135°\}$ on the azimuth plane. The turning limits of the yaw joint in the robot's neck constrain the set of starting angles θ_{neck}. Once the robot orients itself in the first θ_{neck}, we reproduce the first compound sound from the loudspeaker at angle θ_{lspk} and the robot starts tracking the sound source. The trial ends when the sound finishes. Then the robot returns its head to the same angle θ_{neck}, and we reproduce the same compound sound from the next loudspeaker.

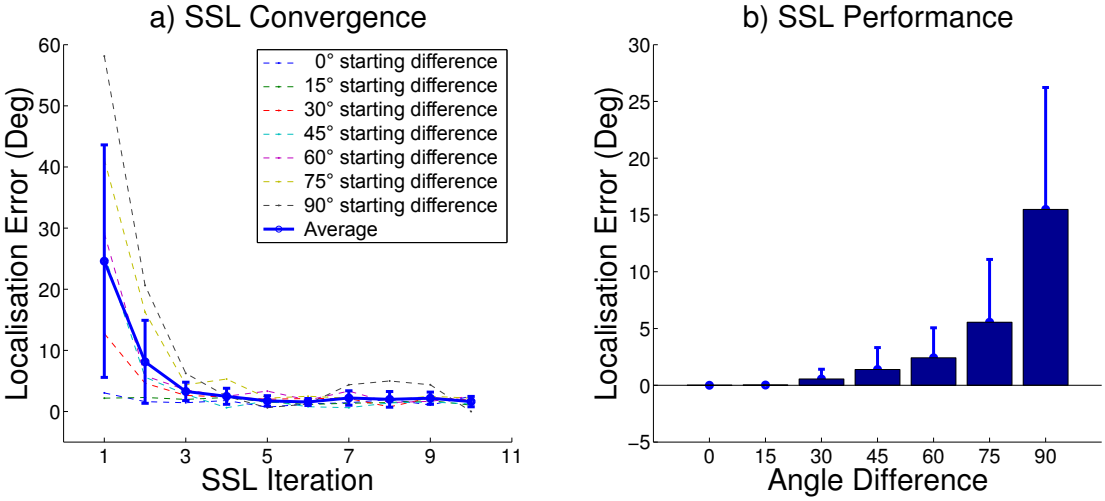


Figure 5.8: Dynamic SSL using the iCub head. **a)** SSL performance in consecutive iterations. The dotted curves display the performance for different angular differences at the beginning of each trial presenting a composed speech sound to the robot. The solid line shows the average of all dotted curves with the bars indicating the standard deviation. Note the small number of steps required for the robot to reach near 0 error, i.e. to face the correct sound source angle. **b)** Accumulated angular error from all iterations in all SSL trials. Note that the accuracy of the SSL system is higher when the angle difference between the sound source and the direction faced by the robot is 0, i.e. when the robot is facing the sound source.

We repeat this procedure until we cover all angles θ_{lspk}. Afterwards, we repeat the same routine over all angles θ_{lspk} for each starting angle θ_{neck}. Finally, we replicate the entire process for each of the 12 compound sounds.

5.3.2 Experimental Results

The results of the dynamic localisation task are summarised in Figure 5.8a for iCub and in Figure 5.9a for Soundman. The figures show the performance of the SSL system in consecutive iterations and from a range of starting angular differences between θ_{neck} and θ_{lspk}, where $\delta_{start} \in \{0°, 15°, \ldots, 90°\}$. The dotted lines in both figures show the average SSL performance of trials with the same starting angular difference δ_{start}. The continuous lines show the average and standard deviations of all starting angular differences δ_{start}. In both figures, we

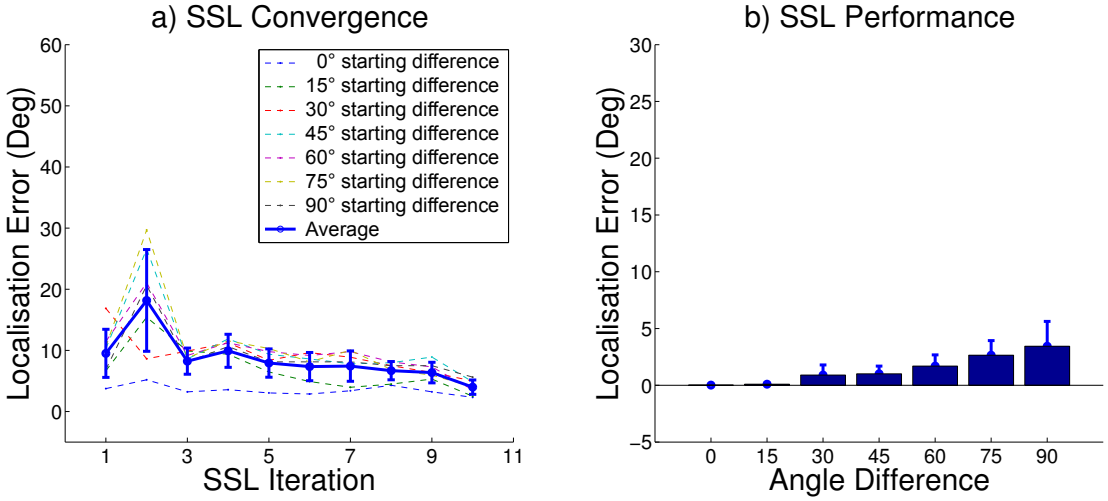


Figure 5.9: Dynamic SSL using the Soundman wooden head. **a)** SSL performance in consecutive iterations. The dotted curves display the performance for different angular differences at the beginning of each trial presenting a composed speech sound to the robot. The solid line shows the average of all dotted curves with the bars indicating the standard deviation. **b)** Accumulated angular error from all iterations in all SSL trials.

can see that the localisation error decreases as δ_{start} decreases from 90° to 0°. The curves show that the system converges to the sound source angle in 3 iterations or less. Afterwards, localisation errors are close to zero with almost no variance. In other words, the SSL system is more robust for localising sounds closer to the front of the head. As localisation errors are smaller in the frontal angles, the SSL system converges to the sound source angle after successive localisation steps. Once the robot is facing the sound source it continues facing that direction, i.e. the SSL system successfully locks the auditory target. These results are consistent with our previous work on static SSL discussed in chapter 5 and with the performance observed in humans (Middlebrooks & Green, 1991).

Figure 5.8b and Figure 5.9b show the angular error accumulated from all SSL iterations. During the experiments many more data points were produced for angles δ_{diff} close to 0°. However, the variance of the accumulated errors also indicates better SSL performance when the sound source is close to the frontal angles. Importantly, this improvement applies to all angles θ_{lspk}. This consistency

in performance shows the robustness of our architecture against the changes in reverberation produced by presenting auditory stimuli from different room locations. Therefore, we conclude that the proposed SSL architecture successfully avoids overfitting to the training data from static sound sources and does not stagnate in poor local minima. It is also important to note that the magnitude of localisation errors is related to the size of the chosen localisation bins (15° of angular granularity). Nevertheless, some preliminary studies show that our system is capable of 1° angular resolution in the frontal 40°. We could access this potential by performing SSL in a continuous space, using the last layer for regression instead of classification. Verifying this hypothesis is part of our following work with the SSL architecture.

5.4 Conclusion

In the *first experiment* in this chapter, about robot speech recognition, we found that using information from SSL can improve the accuracy of speech recognition considerably. As the humanoid platform provides signals from the left and right channels, SSL can indicate how to orient the robot and then select the appropriate channel as input to an ASR system. This approach is in contrast to related approaches that always average the signals from both channels before being using them as input for ASR. Our proposed method is capable of doubling the highest recognition rates at the sentence level when compared to the common averaging method. Interestingly, the performance of the ASR system is not highest when the sound source is facing directly to the microphone in one of the humanoid's ears, but at the angle where the pinna reflects most intensely the sound waves to the microphone. It is possible to measure the magnitude of this improvement by repeating the ASR experiment with the pinnae removed from the heads or even with active pinnae (Kumon & Noda, 2011). A natural extension of the first experiment is to make the robot focus its attention on a single source of information from a possible multitude of concurrent stimuli. It is in this extended scenario, where the input from other sensory modalities comes into play. Vision can be used to disambiguate the location of the speaker in a crowd addressing the robot by observing the orientation of the torso, gaze and lips movement of

every individual detected. Afterwards, this information can be used to perform auditory grouping in time and frequency domains in order to perform speech segregation in noisy environments (Ruggles *et al.*, 2011; Zion Golumbic *et al.*, 2013).

The results of the *second experiment* in this chapter, about acquisition time and source locking, show that the architecture is capable of handling different kinds of reverberation. These results are a significant extension from our previous work in static SSL that support the robustness of the system to the sound dynamics in real-world environments. As another extension considering the dynamics of real-world scenarios, we plan to embed the SSL architecture into a probabilistic framework. In this approach, we can integrate time into the estimation of sound source angles, by using calculations from previous time steps to increase the confidence of the system estimations. This probabilistic model will also benefit from a parallelised version of the MSO and LSO spiking neural layers. In a preliminary GPU implementation, we have already reached 12 times more SSL iterations in the same amount of time than the current CPU version. A fundamental advantage of the neural representation of spatial cues is that we can integrate it directly with visual information for audio-visual spatial attention (Ruesch *et al.*, 2008; Bauer & Wermter, 2013). In this scenario, vision can be used to disambiguate the location of a sound source of interest in a cluttered auditory landscape. As each frequency component generates a spatial hypothesis in our IC model, vision can be used to perform auditory grouping in the time and frequency domains (Zion Golumbic *et al.*, 2013; Lakatos *et al.*, 2013). Furthermore, we can also use vision as a bootstrapping mechanism for training the neural layers in an online fashion. In this way, we can train the entire architecture with an unsupervised learning approach (Nakashima *et al.*, 2002). This unsupervised approach is the main direction of our current research on life-long learning in multimodal speech recognition.

Chapter 6

Conclusions

In this work, we started by reviewing the development of sound localisation methods in the past two decades. Traditional sound source localisation (SSL) techniques can be relatively expensive from a computational point of view, and the required processing units become small enough to be placed on robotic platforms until the 90's decade. This increase in computational power could explain why the first attempts to perform robotic sound localisation appeared around the same years. Such an increase in available computing power in time also justify the evolution of algorithmic approaches in the past decades (Russell & Norvig, 2009). In parallel, neuroscientific theories about perception kept evolving and providing computer scientists with powerful metaphors (Ghahramani, 1995; Kennedy & Dehay, 2012). Many advances in artificial neural networks have taken inspiration from these biological studies to the extent that now we understand better the potential benefits of biomimetic computation. One example of particular interest for us is the inclusion of bioinspired models for interaural time and level differences in the 2000's (Irvine *et al.*, 2001). More specifically, modelling the human auditory pathway with spiking neural networks has proven a robust approach to reverberation and speaker tracking (Dávila-Chacón *et al.*, 2012) as deeper neural architectures appear to be necessary for analysing temporal properties of sound (Christianson *et al.*, 2011; Costa-Faidella *et al.*, 2011).

The latest approaches to sound localisation vary in strengths and weaknesses. Some of the most common challenges include adaptation to changing types of reverberation, segregation of multiple speakers and targets moving around the

array of sensors. Here is where more dynamic approaches can provide an increase in performance. Robots can exploit their capacity to rotate and traverse their environment to increase the confidence in their estimations. Furthermore, embodied embedded cognition has also started to appear in the scene (Metta *et al.*, 2008), for example, humanoid robots can help to produce interaural time (ITDs) and level differences (ILDs) for the estimation of sound sources on the azimuth plane and artificial pinnae have proven to be powerful mechanical filters for the estimation of sound sources on the elevation plane. Finally, state-of-the-art robotic perception points to multimodal integration (Bauer & Wermter, 2013). Different perceptual modalities can benefit from each other, sometimes in fascinating and unexpected manners (Alais & Burr, 2004). Sound source localisation is not the exception and its integration with visual information could improve existing approaches to automatic speech recognition, navigation and scene analysis in daily-life environments (Trifa *et al.*, 2007; Okuno *et al.*, 2007; Liu *et al.*, 2011; Rubruck *et al.*, 2013b).

6.1 Embodied Embedded Cognition and Biomimetic Computation

As stated in our research objectives 1, 2 and 3, our aim has been to improve our understanding of cross-modal integration for acoustic localisation, to understand acoustic localisation from an integrated view of spatial audition at multiple scales and to introduce biological principles into artificial intelligent systems for acoustic localisation. Such guidelines directed our experimental work and led us to interesting conclusions in our three main research objectives.

Our *first objective*, was to increase our understanding of the influence of humanoid embodiment on bottom-up cognitive tasks for sound perception (Koch, 1993), such as static and dynamic SSL. For this reason, we do not focus our research on the design of an SSL or an ASR system that performs better than other systems designed in the past, but our primary objective is to improve on our understanding of the influence of the body on bottom-up cognitive tasks. If

the best interface for a human, is another human, we should also exploit the computational advantages that this embodiment brings *for free*, if any (Asada *et al.*, 2001; Kanda *et al.*, 2004). With the iCub, we can approach the physiognomy of humans and measure the influence that it has on our models of the auditory system. Here is where we aim to provide a clear answer based on our results, and hence, we compare the effect of the embodiment of the iCub against the effect of the silent dummy head. Ultimately, the idea is to find the advantages for top-down cognition when using biomimetic models of bottom-up cognition. Once the behaviour of the robot corresponds to the behaviour of animals, we can observe the activity of the neural models under new conditions and produce new hypotheses to guide further studies in biological systems.

Our *second objective*, was to increase our understanding about the influence of embodiment on top-down cognitive tasks (Koch, 1993; Zhao *et al.*, 2018) like ASR, when using biomimetic models of bottom-up cognition like SSL. This objective led us to measure the capacity of different robotic platforms to produce the necessary spatial cues for SSL under increasingly difficult conditions. Now, building on the premise that we can integrate different auditory cues in the same way as multisensory information, we focused on the design of an SSL system that could take advantage of the spatial cues produced by the interaction between sound waves and the embodiment of humanoid robots. We represent such cues with biomimetic models of regions in the auditory pathway that convert them into spatial representations embedded in the topology of neural populations. Our architecture then integrates these cues with a Bayesian model of the inferior colliculus (IC) that performs dimensionality reduction and finally a multilayer perceptron with a probabilistic layer converted the output of the IC into commands for motor control of the robot (Rokni & Sompolinsky, 2012). This architecture proved to be robust to high levels of noise and led us to explore its potential for supporting automatic speech recognition (ASR) in robotic platforms with ego-noise. Our last experiments show that such humanoid platforms improved the performance of ASR considerably when they adapt their orientation to increase the signal-to-noise ratio of the speech signal. Interestingly, this angle lies around 45° from the sagittal plane, rather than when facing the speaker. The following subsection dives more deeply into the technical lessons learned in our journey.

Our *third objective*, was to close the loop by using the experimental results obtained with artificial systems to guide further research in natural systems. One of the first observations that we had on the activation patterns of the spiking neural networks in our system, was that both spatial cues are complementary as they represent information in opposite ends of the auditive spectrum. More specifically, the importance of integrating ITDs and ILDs can be understood further by observing the overlap of excitatory connections from the model of the medial superior olive (MSO), and excitatory and inhibitory connections from the model of the lateral superior olive (LSO). On the one hand, neurones in the MSO model have informative activity in all frequencies but also potentially misleading activity in higher frequencies. On the other hand, neurones in the LSO model have informative activity only in higher frequencies. For this reason, LSO excitatory connections to the IC reinforce useful activity from high frequencies in the MSO, while LSO inhibitory connections to the IC remove the misleading activity from high frequencies in the MSO (Dávila-Chacón *et al.*, 2012; de Queiroz *et al.*, 2006). An interesting application of the activation patterns that emerge in our models of the MSO, LSO and IC, could be to predict the neural activity in the respective layers in the mammalian brainstem when presenting different types of stimuli. Having a model that correctly predicts such activation patterns in natural systems could potentially speed up the development of medical applications.

Having these learnings from our time researching the impact of embodied embedded cognition and biomimetic computation on automatic speech recognition, we feel confident that the iCub platform is capable of representing spatial cues close enough to human bodies, independently of the difference in internal structure, in materials and the high levels of ego-noise. This conclusion arises from the similarity of activation patterns in the spiking neural networks used to represent different spatial cues, and the localisation accuracy achieved by the overall system. What can be done next to increment our understanding of the higher levels of cognition in the human auditory system?

6.2 Future Work

Continuous life-long learning in neural models is an open topic where SSL can offer useful insights from our current knowledge in natural systems (Block & Bastian, 2011; Phillips *et al.*, 2011; Wagner & Dobkins, 2011). It is an ongoing research topic to understand how top-down connections in the auditory pathway influence the neural activity in the very first stages of processing, i.e., all the way to the cochlea (Nodal *et al.*, 2010). What is known, nevertheless, is that the auditory nerve also has neural pathways descending from the medial olivocochlear system to the cochlea (Andéol *et al.*, 2011). These projections are known to affect the neural representations of direction-dependent spectral features, which are crucial for accurate localisation in the elevation plane and front/back disambiguation. We found that many of the state-of-the-art methods for SSL consist of deep neural networks that lack descending connections and therefore have ample possibilities for research, such as remapping relative to motion (Teramoto *et al.*, 2012) and development (Sinapov *et al.*, 2011).

Along the chapters detailing our experiments we consistently suggest the potential benefits of extending our current SSL architecture with multimodal integration (Battaglia *et al.*, 2003; Bauer *et al.*, 2012, to appear). This work has already been started by Bauer & Wermter (2013), and shows promising extensions over preceeding approaches to SSL (Kim *et al.*, 2007, 2011). In this direction, it is possible to integrate input from vision to define the required processing to handle the type of reverberation present in a given environment (Liu & Yang, 2014). Similarly, ASR could update the relevant Markov model to fit with statistical correlation for specific rooms, people and other cues from vision (Martinson & Schultz, 2007; Harrison & De Kamps, 2011). An exciting extension of this research could be to test generative adversarial networks (Goodfellow *et al.*, 2014) for the dynamic inference of masks for SSL and domain dependent language models for ASR (Julian *et al.*, 2017; Liu *et al.*, 2017).

Also, due to the sequential nature of sound, the task of SSL can be seen learned by a reward function dependent on feedback from vision. Reinforcement learning (RL) architectures (Sutton & Barto, 1998) have made considerable progress in recent years to cope with large state and action spaces (Silver *et al.*, 2014; Lillicrap

et al., 2015). When using top-down attention mechanisms that follow bottom-up salient features such as primary colours, motion and sound intensity, RL could act as a bootstrap mechanism to guide the development of spatial maps. Such maps can represent auditory cues with a topology corresponding to spatial locations around a robotic platform and would help to test theories of neural development in children (Greene & Oliva, 2009; Boes *et al.*, 2012). Furthermore, unsupervised methods are particularly relevant for life-long learning and several works offer a sound starting point to extend our approach (Nakashima *et al.*, 2002; Kim & Choi, 2009; Kitani *et al.*, 2012).

As noted so far, there is ample potential for the further development of our biomimetic SSL models and its extension into a multimodal integration system. Particularly, their implementation with asynchronous computation would allow us to test their performance in real time (Igarashi *et al.*, 2011). Our group has all the necessary equipment and testing facilities (Bauer *et al.*, 2012) to monitor such experiments under tight control, and also to compare the results obtained with the computational models against the behaviour of human subjects. We are in an inspiring time for exploiting the new possibilities brought by the availability of large amounts of computing power and novel algorithmic approaches. The current work has made a modest contribution in our understanding of the interaction between embodied embedded cognition and biomimetic computation for humanoid robots, and now it is the right time to build further towards a more general architecture of human cognition.

Appendices

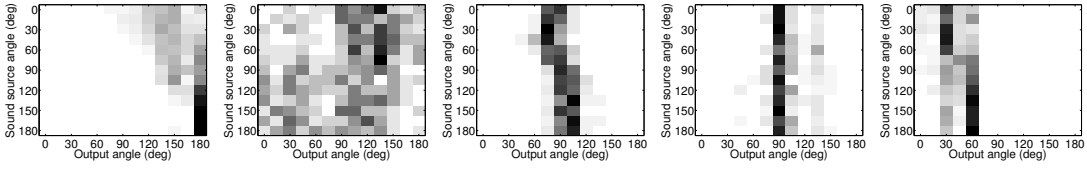
Appendix A

Supplementary Experimental Results

The global results from all experiments carried out in this chapter are shown in the following subsections. The best results are discussed in section 4.3.

A.0.1 Winner Takes All

This subsection contains the confusion matrices and performance tables when using Winner Takes All (WTA) for classification. The results after training with white noise and testing with speech can be seen in Figure A.1 and Table A.1. The results after training with speech and testing with white noise can be seen in Figure A.2 and Table A.2.

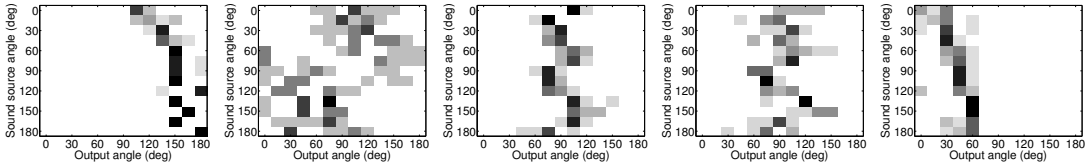


(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.1: WTA: Training with White Noise / Testing with Speech.

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.05	0.26	0.61	0.59	0.08
MSO	0.11	0.60	0.51	0.52	0.19
LSO	0.02	0.11	0.64	0.60	0.03
MSO-LSO	0.07	0.36	0.62	0.60	0.12
Bayes	0.05	0.19	0.72	0.68	0.07

Table A.1: WTA: Training with White Noise / Testing with Speech.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

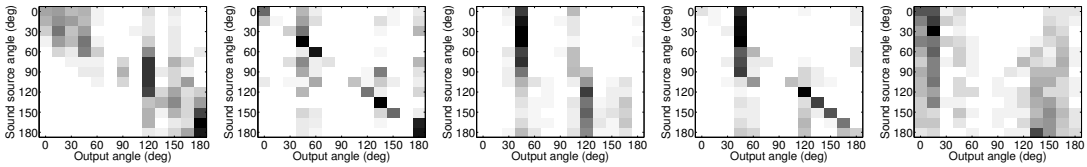
Figure A.2: WTA: Training with Speech / Testing with White Noise.

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.06	0.25	0.68	0.65	0.10
MSO	0.06	0.44	0.50	0.49	0.11
LSO	0.08	0.41	0.62	0.61	0.13
MSO-LSO	0.06	0.44	0.50	0.49	0.11
Bayes	0.01	0.05	0.71	0.67	0.02

Table A.2: WTA: Training with Speech / Testing with White Noise.

A.0.2 K Nearest Neighbours

This subsection contains the confusion matrices and performance tables when using K Nearest Neighbours (KNN) for classification. The results after training with white noise and testing with speech can be seen in Figure A.3 and Table A.3. The results after training with speech and testing with white noise can be seen in Figure A.4 and Table A.4.

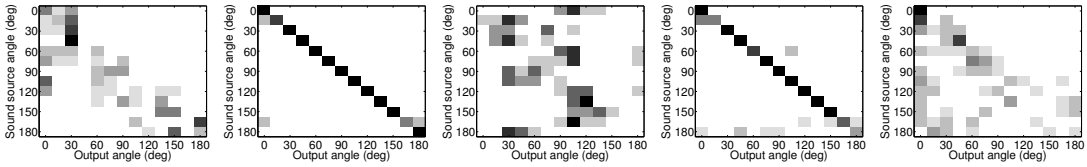


(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.3: KNN: Training with White Noise / Testing with Speech.

	Precision	Recall	True Neg	Accuracy	F-Measure
CCR	0.26	0.79	0.58	0.61	0.39
MSO	0.69	0.93	0.78	0.83	0.79
LSO	0.11	0.57	0.55	0.55	0.18
MSO-LSO	0.63	0.92	0.74	0.80	0.75
Bayes	0.10	0.56	0.53	0.53	0.17

Table A.3: KNN: Training with White Noise / Testing with Speech.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.4: KNN: Training with Speech / Testing with White Noise.

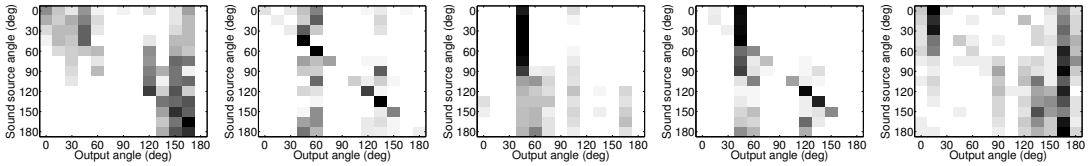
	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.21	0.75	0.59	0.61	0.33
MSO	0.96	1.00	0.96	0.98	0.98
LSO	0.15	0.65	0.54	0.55	0.25
MSO-LSO	0.89	0.99	0.90	0.94	0.93
Bayes	0.18	0.75	0.53	0.56	0.29

Table A.4: KNN: Training with Speech / Testing with White Noise.

A.0.3 Learning Vector Quantisation

This subsection contains the confusion matrices and performance tables when using Learning Vector Quantisation (LVQ) for classification. The results after

training with white noise and testing with speech can be seen in Figure A.5 and Table A.5. The results after training with speech and testing with white noise can be seen in Figure A.6 and Table A.6.

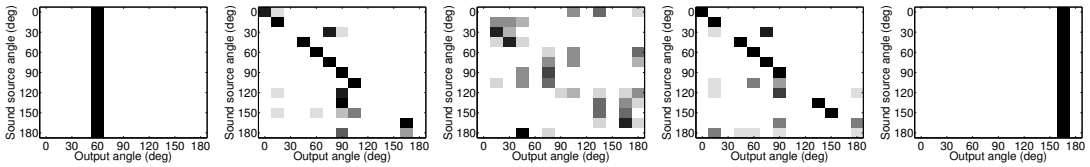


(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.5: LVQ: Training with White Noise / Testing with Speech.

	Precision	Recall	True Neg	Accuracy	F-Measure
CCR	0.22	0.73	0.61	0.63	0.34
MSO	0.67	0.93	0.76	0.82	0.78
LSO	0.08	0.37	0.65	0.63	0.13
MSO-LSO	0.68	0.90	0.81	0.84	0.77
Bayes	0.22	0.75	0.54	0.57	0.34

Table A.5: LVQ: Training with White Noise / Testing with Speech.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

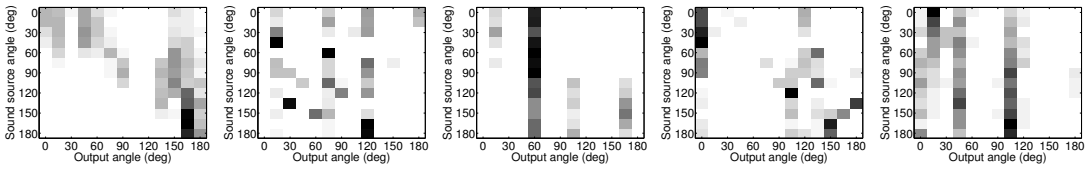
Figure A.6: LVQ: Training with Speech / Testing with White Noise.

A.0.4 Self Organising Map

This subsection contains the confusion matrices and performance tables when using Self Organising Map (SOM) for classification. The results after training with white noise and testing with speech can be seen in Figure A.7 and Table A.7. The results after training with speech and testing with white noise can be seen in Figure A.8 and Table A.8.

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.08	0.08	0.92	0.86	0.08
MSO	0.73	0.94	0.85	0.88	0.82
LSO	0.16	0.62	0.62	0.62	0.26
MSO-LSO	0.68	0.94	0.81	0.85	0.79
Bayes	0.08	0.08	0.92	0.86	0.08

Table A.6: LVQ: Training with Speech / Testing with White Noise.

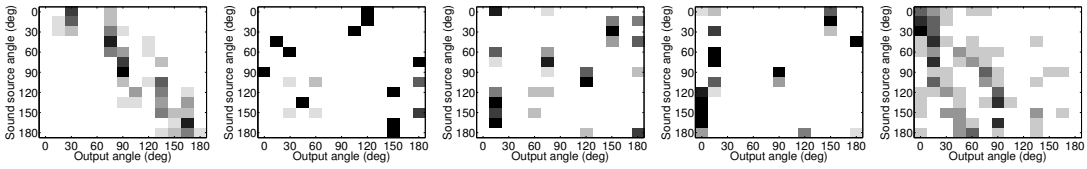


(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.7: SOM: Training with White Noise / Testing with Speech.

	Precision	Recall	True Neg	Accuracy	F-Measure
CCR	0.19	0.71	0.58	0.59	0.30
MSO	0.02	0.14	0.55	0.53	0.03
LSO	0.17	0.44	0.79	0.75	0.25
MSO-LSO	0.12	0.59	0.53	0.54	0.19
Bayes	0.09	0.47	0.60	0.59	0.16

Table A.7: SOM: Training with White Noise / Testing with Speech.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.8: SOM: Training with Speech / Testing with White Noise.

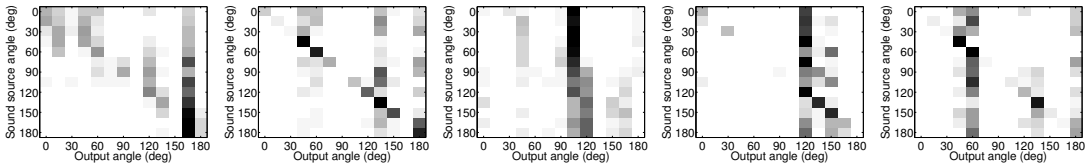
A.0.5 Multilayer Perceptron

This subsection contains the confusion matrices and performance tables when using Multilayer Perceptron (MLP) for classification. The results after training

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.30	0.79	0.64	0.66	0.43
MSO	0.00	0.00	0.57	0.54	NaN
LSO	0.14	0.47	0.70	0.68	0.22
MSO-LSO	0.14	0.47	0.70	0.68	0.22
Bayes	0.05	0.37	0.54	0.53	0.09

Table A.8: SOM: Training with Speech / Testing with White Noise.

with white noise and testing with speech can be seen in Figure A.9 and Table A.9. The results after training with speech and testing with white noise can be seen in Figure A.10 and Table A.10.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

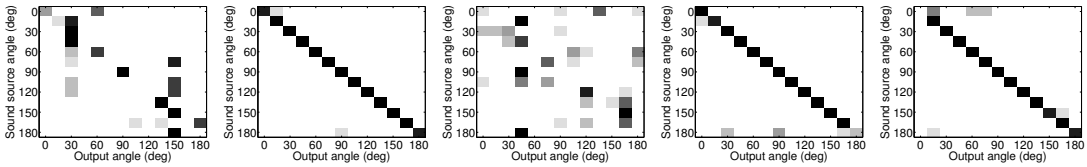
Figure A.9: MLP: Training with White Noise / Testing with Speech.

	Precision	Recall	True Neg	Accuracy	F-Measure
CCR	0.36	0.83	0.63	0.67	0.50
MSO	0.75	0.94	0.82	0.86	0.83
LSO	0.08	0.48	0.54	0.54	0.14
MSO-LSO	0.62	0.85	0.82	0.83	0.72
Bayes	0.71	0.92	0.79	0.84	0.80

Table A.9: MLP: Training with White Noise / Testing with Speech.

A.0.6 Radial Basis Functions

This subsection contains the confusion matrices and performance tables when using Radial Basis Functions (RBF) for classification. The results after training with white noise and testing with speech can be seen in Figure A.11 and Table



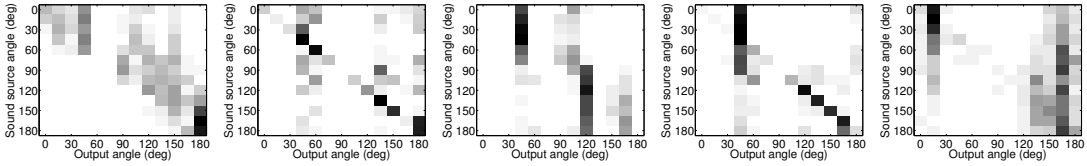
(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.10: MLP: Training with Speech / Testing with White Noise.

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.55	0.89	0.75	0.79	0.68
MSO	0.98	1.00	0.98	0.99	0.99
LSO	0.21	0.76	0.56	0.59	0.33
MSO-LSO	0.95	0.99	0.95	0.97	0.97
Bayes	0.93	0.99	0.94	0.96	0.96

Table A.10: MLP: Training with Speech / Testing with White Noise.

A.11. The results after training with speech and testing with white noise can be seen in Figure A.12 and Table A.12.

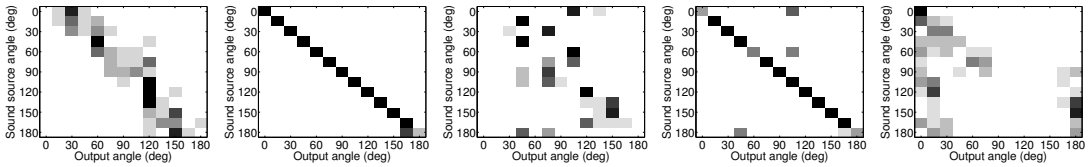


(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.11: RBF: Training with White Noise / Testing with Speech.

	Precision	Recall	True Neg	Accuracy	F-Measure
CCR	0.24	0.76	0.59	0.62	0.37
MSO	0.71	0.94	0.78	0.84	0.81
LSO	0.17	0.58	0.68	0.67	0.26
MSO-LSO	0.69	0.93	0.77	0.83	0.79
Bayes	0.33	0.83	0.58	0.63	0.48

Table A.11: RBF: Training with White Noise / Testing with Speech.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

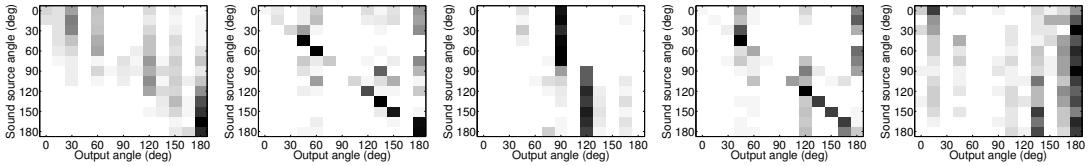
Figure A.12: RBF: Training with Speech / Testing with White Noise.

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.20	0.75	0.56	0.58	0.32
MSO	0.97	1.00	0.97	0.98	0.98
LSO	0.42	0.83	0.70	0.73	0.56
MSO-LSO	0.93	0.99	0.93	0.96	0.96
Bayes	0.27	0.73	0.68	0.68	0.40

Table A.12: RBF: Training with Speech / Testing with White Noise.

A.0.7 Clustering with K-Means and Classification with Multilayer Perceptron

This subsection contains the confusion matrices and performance tables when clustering with K-Means and classifying with Multilayer Perceptron (KM + MLP). The results after training with white noise and testing with speech can be seen in Figure A.13 and Table A.13. The results after training with speech and testing with white noise can be seen in Figure A.14 and Table A.14.

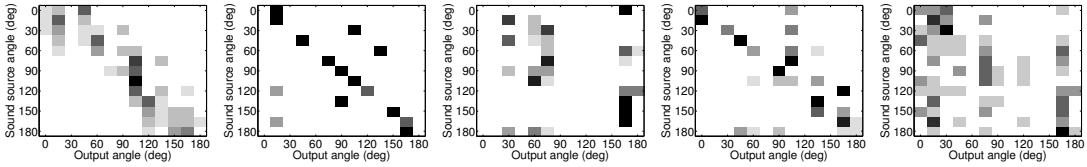


(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.13: KM + MLP: Training with White Noise / Testing with Speech.

	Precision	Recall	True Neg	Accuracy	F-Measure
CCR	0.18	0.72	0.55	0.57	0.29
MSO	0.70	0.93	0.79	0.84	0.80
LSO	0.13	0.48	0.66	0.65	0.20
MSO-LSO	0.69	0.93	0.77	0.83	0.79
Bayes	0.15	0.61	0.59	0.59	0.24

Table A.13: KM + MLP: Training with White Noise / Testing with Speech.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

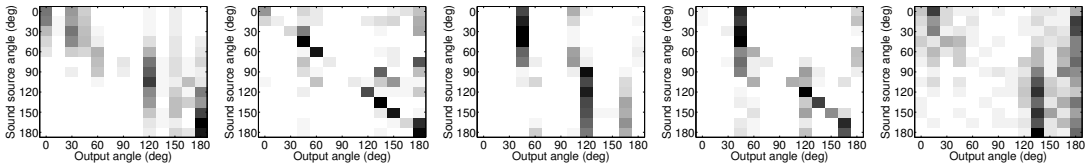
Figure A.14: KM + MLP: Training with Speech / Testing with White Noise.

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.20	0.75	0.56	0.59	0.32
MSO	0.64	0.93	0.79	0.83	0.76
LSO	0.10	0.36	0.73	0.70	0.16
MSO-LSO	0.53	0.89	0.75	0.78	0.67
Bayes	0.20	0.68	0.63	0.63	0.31

Table A.14: KM + MLP: Training with Speech / Testing with White Noise.

A.0.8 Clustering with K-Means and Classification with Radial Basis Functions

This subsection contains the confusion matrices and performance tables when using clustering with K-Means and classifying with Radial Basis Functions (KM + RBF). The results after training with white noise and testing with speech can be seen in Figure A.15 and Table A.15. The results after training with speech and testing with white noise can be seen in Figure A.16 and Table A.16.

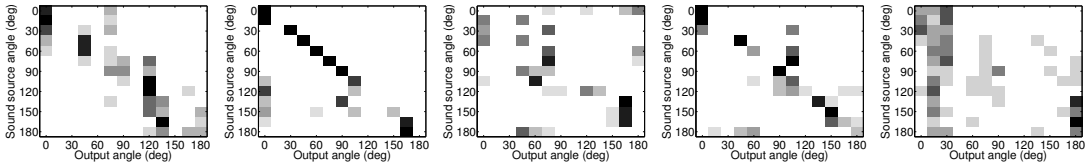


(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.15: KM + RBF: Training with White Noise / Testing with Speech.

	Precision	Recall	True Neg	Accuracy	F-Measure
CCR	0.24	0.76	0.59	0.62	0.36
MSO	0.71	0.94	0.78	0.84	0.81
LSO	0.20	0.66	0.66	0.66	0.31
MSO-LSO	0.59	0.91	0.73	0.78	0.71
Bayes	0.16	0.68	0.52	0.54	0.26

Table A.15: KM + RBF: Training with White Noise / Testing with Speech.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

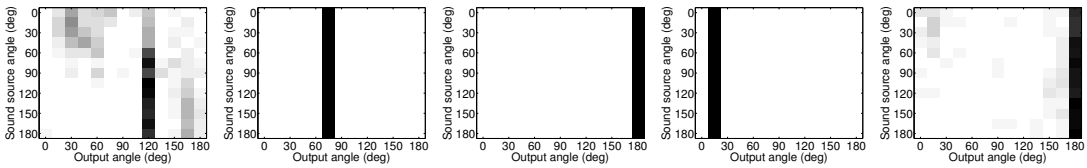
Figure A.16: KM + RBF: Training with Speech / Testing with White Noise.

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.27	0.74	0.68	0.68	0.39
MSO	0.75	0.94	0.85	0.88	0.84
LSO	0.21	0.72	0.61	0.62	0.33
MSO-LSO	0.57	0.92	0.72	0.78	0.70
Bayes	0.12	0.56	0.58	0.58	0.20

Table A.16: KM + RBF: Training with Speech / Testing with White Noise.

A.0.9 Clustering with Self Organising Map and Classification with Multilayer Perceptron

This subsection contains the confusion matrices and performance tables when clustering with Self Organising Map and classifying with Multilayer Perceptron (SOM + MLP). The results after training with white noise and testing with speech can be seen in Figure A.17 and Table A.17. The results after training with speech and testing with white noise can be seen in Figure A.18 and Table A.18.

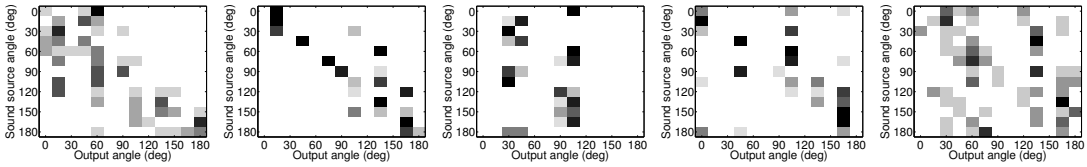


(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.17: SOM + MLP: Training with White Noise / Testing with Speech.

	Precision	Recall	True Neg	Accuracy	F-Measure
CCR	0.20	0.74	0.54	0.57	0.32
MSO	0.08	0.08	0.92	0.86	0.08
LSO	0.08	0.08	0.92	0.86	0.08
MSO-LSO	0.08	0.08	0.92	0.86	0.08
Bayes	0.12	0.56	0.58	0.58	0.20

Table A.17: SOM + MLP: Training with White Noise / Testing with Speech.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

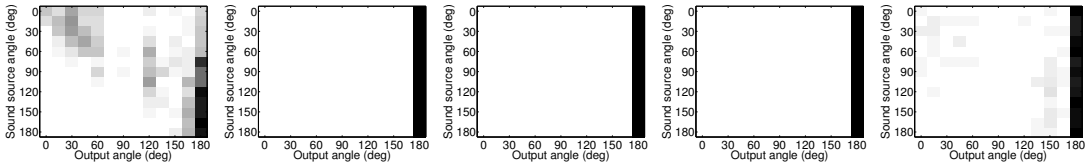
Figure A.18: SOM + MLP: Training with Speech / Testing with White Noise.

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.22	0.76	0.56	0.59	0.34
MSO	0.66	0.91	0.82	0.84	0.76
LSO	0.15	0.39	0.78	0.74	0.21
MSO-LSO	0.41	0.76	0.77	0.77	0.53
Bayes	0.15	0.64	0.56	0.57	0.25

Table A.18: SOM + MLP: Training with Speech / Testing with White Noise.

A.0.10 Clustering with Self Organising Map and Classification with Radial Basis Functions

This subsection contains the confusion matrices and performance tables when clustering with Self Organising Map and classifying with Radial Basis Functions (SOM + RBF). The results after training with white noise and testing with speech can be seen in Figure A.19 and Table A.19. The results after training with speech and testing with white noise can be seen in Figure A.20 and Table A.20.

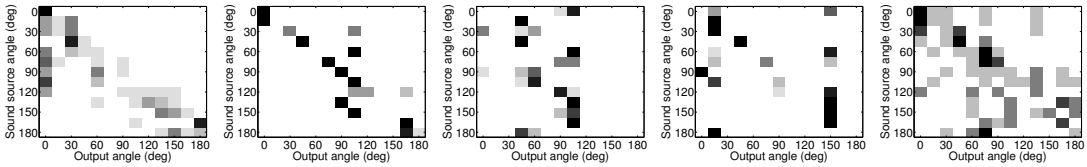


(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.19: SOM + RBF: Training with White Noise / Testing with Speech.

	Precision	Recall	True Neg	Accuracy	F-Measure
CCR	0.27	0.77	0.60	0.63	0.40
MSO	0.08	0.08	0.92	0.86	0.08
LSO	0.08	0.08	0.92	0.86	0.08
MSO-LSO	0.08	0.08	0.92	0.86	0.08
Bayes	0.12	0.56	0.58	0.58	0.20

Table A.19: SOM + RBF: Training with White Noise / Testing with Speech.



(a) CCR **(b)** MSO **(c)** LSO **(d)** MSO-LSO **(e)** Bayes

Figure A.20: SOM + RBF: Training with Speech / Testing with White Noise.

	Precision	Recall	True Negative	Accuracy	F-Measure
CCR	0.23	0.78	0.57	0.60	0.36
MSO	0.75	0.94	0.85	0.88	0.83
LSO	0.06	0.23	0.72	0.68	0.10
MSO-LSO	0.40	0.76	0.77	0.77	0.53
Bayes	0.25	0.80	0.56	0.59	0.39

Table A.20: SOM + RBF: Training with Speech / Testing with White Noise.

A. SUPPLEMENTARY EXPERIMENTAL RESULTS

References

Agnes, E.J., Jr, R.E. & Brunnet, L.G. (2012). Associative Memory in Neuronal Networks of Spiking Neurons: Architecture and Storage Analysis. 145–152. 25

Alais, D. & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, **14**, 257–262. 94

Amari, S.i. (2013). Dreaming of mathematical neuroscience for half a century. *Neural Networks*, **37**, 48–51. 25

Andéol, G., Guillaume, A., Micheyl, C., Savel, S., Pellieux, L. & Moulin, A. (2011). Auditory efferents facilitate sound localization in noise in humans. *Journal of Neuroscience*, **31**, 6759–6763. 97

Andersson, S.B., Handzel, a.a., Shah, V. & Krishnaprasad, P.S. (2004). Robot phonotaxis with dynamic sound-source localization. *IEEE International Conference on Robotics and Automation 2004 Proceedings ICRA 04 2004*, **5**, 4833–4838. 2, 10

Asada, M., MacDorman, K.F., Ishiguro, H. & Kuniyoshi, Y. (2001). Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, **37**, 185–193. 80, 95

Asano, F., Goto, M., Itou, K. & Asoh, H. (2001). Real-time sound source localization and separation system and its application to automatic speech recognition. In *INTERSPEECH*, 1013–1016. 38

Ashida, G. & Carr, C.E. (2011). Sound localization: Jeffress and beyond. *Current Opinion in Neurobiology*, **21**, 745–751. 62

REFERENCES

Asono, F., Asoh, H. & Matsui, T. (1999). Sound source localization and signal separation for office robot "Jijo-2". In *Multisensory Fusion and Integration for Intelligent Systems*, 243–248, IEEE. 22

Atencio, C.A., Sharpee, T.O. & Schreiner, C.E. (2012). Receptive field dimensionality increases from the auditory midbrain to cortex. *Journal of Neurophysiology*, **107**, 2594–2603. 10

Baken, R.J. & Orlikoff, R.F. (2000). *Clinical measurement of speech and voice*. Cengage Learning. 49

Barker, J.P., Cooke, M.P. & Ellis, D.P.W. (2005). Decoding speech in the presence of other sources. *Speech Communication*, **45**, 5–25. 79

Barrès, V., Simons, A. & Arbib, M. (2013). Synthetic event-related potentials: a computational bridge between neurolinguistic models and experiments. *Neural networks: the official journal of the International Neural Network Society*, **37**, 66–92. 13

Battaglia, P.W., Jacobs, R.A. & Aslin, R.N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, **20**, 1391–1397. 97

Bauer, J. & Wermter, S. (2013). Self-organized neural learning of statistical inference from high-dimensional data. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence, IJCAI*, 1226–1232, AAAI Press. 92, 94, 97

Bauer, J., Davila-Chacon, J., Strahl, E. & Wermter, S. (2012). Smoke and mirrors—virtual realities for sensor fusion experiments in biomimetic robotics. In *International Conference on Multisensor Fusion and Integration, MFI*, 114–119, IEEE. 14, 38, 57, 78, 79, 80, 98

Bauer, J., Weber, C. & Wermter, S. (2012, to appear). A SOM-based model for multi-sensory integration in the superior colliculus. In *Proceedings of the International Joint Conference on Neural Networks (2012: Brisbane, Australia)*, IEEE. 97

Beira, R., Lopes, M., Praga, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchi, F. & Saltarén, R. (2006). Design of the robot-cub (iCub) head. In *International Conference on Robotics and Automation, ICRA*, 94–100, IEEE. 13, 28, 58, 80

Benesty, J., Sondhi, M. & Huang, Y. (2007). *Springer handbook of speech processing*. Springer. 60

Benoit, C., Martin, J.C., Pelachaud, C., Schomaker, L. & Suhm, B. (2000). Audio-visual and multimodal speech systems. *Handbook of Standards and Resources for Spoken Language Systems-Supplement*, **500**. 14

Bentley, J.L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, **18**, 509–517. 67

Berglund, E. & Sitte, J. (2005). Sound source localisation through active audition. *2005 IEEERSJ International Conference on Intelligent Robots and Systems*, 653–658. 19

Besson, P., Bourdin, C. & Bringoux, L. (2011). A comprehensive model of audiovisual perception: both percept and temporal dynamics. *PloS one*, **6**, e23811. 19

Bhadkamkar, N. (1994). Binaural source localizer chip using subthreshold analog CMOS. In *International Conference on Neural Networks*, vol. 3, 1866–1870, IEEE. 20

Biologie, V.F. (2007). Computer simulation of chopper neurons: intrinsic oscillations and temporal processing in the auditory system. *Neuroanatomy*. 9

Bisani, M. & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, **50**, 434–451. 84

Blauert, J. (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT press, Cambridge. 3

Block, H.J. & Bastian, A.J. (2011). Sensory weighting and realignment: independent compensatory processes. *Journal of Neurophysiology*, **106**, 59–70. 97

Boes, M., Oldoni, D., Coensel, B.D. & Botteldooren, D. (2012). Attention-Driven Auditory Stream Segregation using a SOM coupled with an Excitatory-Inhibitory ANN. 10–15. 98

Bowers, J.S. (2009). On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychological Review*, **116**, 220–251. 2

REFERENCES

Brand, A., Behrend, O., Marquardt, T., McAlpine, D. & Grothe, B. (2002). Precise inhibition is essential for microsecond interaural time difference coding. *Nature*, **417**, 543–547. 10

Brette, R. (2012). Spiking models for level-invariant encoding. *Frontiers in Computational Neuroscience*, **5**, 63. 4, 10

Chan, V.Y.S., Jin, C.T. & Van Schaik, A. (2010). Adaptive Sound Localization with a Silicon Cochlea Pair. *Frontiers in neuroscience*, **4**, 11. 25

Chan, V.Y.S., Jin, C.T. & Van Schaik, A. (2012). Neuromorphic audio-visual sensor fusion on a sound-localizing robot. *Frontiers in Neuroscience*, **6**, 1–9. 25

Chase, S.M. & Young, E.D. (2008). Cues for sound localization are encoded in multiple aspects of spike trains in the inferior colliculus. *Journal of neurophysiology*, **99**, 1672–1682. 10

Choudhary, S., Sloan, S., Fok, S., Neckar, A., Trautmann, E., Gao, P., Stewart, T., Eliasmith, C. & Boahen, K. (2012). Silicon Neurons That Compute. 121–128. 25

Christianson, G., Sahani, M. & Linden, J. (2011). Temporal response properties in auditory cortex are depth-dependent. *Journal of Neuroscience*, **31**, 12837–12848. 93

Cobos, M., Marti, A. & Lopez, J.J. (2011). A modified srp-phat functional for robust real-time sound source localization with scalable spatial sampling. *Signal Processing Letters, IEEE*, **18**, 71–74. 25

Cong-qing, L., Fang, W., Shi-jie, D., Li-xin, S., He, H. & Li-ying, S. (2009). A novel method of binaural sound localization based on dominant frequency separation. In *International Cong. on Image and Signal Processing, CISP*, 1–4, IEEE. 39

Costa-Faidella, J., Baldeweg, T., Grimm, S. & Escera, C. (2011). Interactions between "what" and "when" in the auditory system: temporal predictability enhances repetition suppression. *Journal of Neuroscience*, **31**, 18590–18597. 93

Cover, T. & Hart, P. (1967). Nearest neighbor pattern classification. *Transactions on Information Theory*, **13**, 21–27. 67

Davila-Chacon, J., Magg, S., Liu, J. & Wermter, S. (2013). Neural and statistical processing of spatial cues for sound source localisation. In *International Joint Conference on Neural Networks, IJCNN*, IEEE. 27, 28, 36, 84

de Queiroz, M., de Berredo, R. & de Padua Braga, A. (2006). Reinforcement learning of a simple control task using the spike response model. *Neurocomputing*, **70**, 14–20. 96

Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., Kubo, Y., Souden, M., Hahm, S.j. & Nakamura, A. (2011). Speech recognition in the presence of highly non-stationary noise based on spatial , spectral and temporal speech / noise modeling combined with dynamic variance adaptation. 12–17. 2

Deleforge, A. & Horaud, R. (2012). The cocktail party robot: Sound source separation and localisation with an active binaural head. In *Proceedings of the International Conference on Human-Robot Interaction*, 431–438, ACM/IEEE. 39

Devore, S., Ihlefeld, A., Hancock, K., Shinn-Cunningham, B. & Delgutte, B. (2009). Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain. *Neuron*, **62**, 123–134. 40

Dosher, B.A. & Lu, Z.L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 13988–13993. 2

Dávila-Chacón, J., Heinrich, S., Liu, J. & Wermter, S. (2012). Biomimetic binaural sound source localisation with ego-noise cancellation. In *Proceedings of the International Conference on Artificial Neural Networks (2012: Lausanne, Swiss)*, Lecture Notes in Computer Science, Springer. 27, 62, 71, 76, 93, 96

Ernst, M.O. & Bülthoff, H.H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, **8**, 162–169. 2

Escabi, M.a. & Schreiner, C.E. (2002). Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *Journal of Neuroscience*, **22**, 4114–4131. 10

Even, J., Heracleous, P., Ishi, C. & Hagita, N. (2011). Multi-modal front-end for speaker activity detection in small meetings. In *International Conference on Intelligent Robots and Systems, IROS*, 536–541, IEEE. 55

Famulare, M. & Fairhall, A. (2010). Feature selection in simple neurons: how coding depends on spiking dynamics. *Neural computation*, **22**, 581–598. 13

Finger, H. & Liu, S.C. (2011). Estimating the location of a sound source with a spike-timing localization algorithm. *2011 IEEE International Symposium of Circuits and Systems ISCAS*, 2461–2464. 13

Finger, H., Liu, S.C.L.S.C., Ruvolo, P. & Movellan, J.R. (2010). Approaches and databases for online calibration of binaural sound localization for robotic heads. *Intelligent Robots and Systems IROS 2010 IEEERSJ International Conference on*, 4340–4345. 81

Fischer, B.J. & Peña, J.L. (2011). Owl's behavior and neural representation predicted by Bayesian inference. *Nature Neuroscience*, **14**, 1061–1066. 14

Fontaine, B. & Brette, R. (2011). Neural Development of Binaural Tuning through Hebbian Learning Predicts Frequency-Dependent Best Delays. *Journal of Neuroscience*, **31**, 11692–11696. 14

Fréchette, M., Létourneau, D., Valin, J. & Michaud, F. (2012). Integration of sound source localization and separation to improve dialogue management on a robot. In *International Conference on Intelligent Robots and Systems, IROS*, 2358–2363, IEEE. 38

Futagi, D. & Kitano, K. (2012). A Biologically Realizable Bayesian Computation. 247–254. 27

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G. & Pallett, D.S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, **93**, 27403. 37, 84

Geng, Y.G.Y., Jung, J.J.J. & Seol, D.S.D. (2008). Sound-source localization system based on neural network for mobile robots. *2008 IEEE International Joint Conference on Neural Networks IEEE World Congress on Computational Intelligence*, 3125–3129. 25

Ghahramani, Z. (1995). *Computation and Psychophysics of Sensorimotor Integration*. Ph.D. thesis, Massachusetts Institute of Technology. 93

Glackin, B., Wall, J.a., McGinnity, T.M., Maguire, L.P. & McDaid, L.J. (2010). A Spiking Neural Network Model of the Medial Superior Olive Using Spike Timing Dependent Plasticity for Sound Localization. *Frontiers in computational neuroscience*, **4**, 16. 14

Glendenning, K. & Masterton, R. (1983). Acoustic chiasm: Efferent projections of the lateral superior olive. *The Journal of Neuroscience*, **3**, 1521–1537. 10

Goertzel, B., Lian, R., Arel, I., De Garis, H. & Chen, S. (2010). A world survey of artificial brain projects, Part II: Biologically inspired cognitive architectures. *Neurocomputing*, **74**, 30–49. 14

Golumbic, E.M.Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z. *et al.* (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, **77**, 980–991. 2

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680. 97

Goodman, D.F.M. & Brette, R. (2010). Spike-timing-based computation in sound localization. *PLoS computational biology*, **6**, e1000993. 4

Goodrich, M.A. & Schultz, A.C. (2007). Human–robot interaction: A survey. *Human–Computer Interaction*, **1**, 203–275. 88

Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J. & Maisonnier, B. (2009). Mechatronic design of NAO humanoid. In *International Conference on Robotics and Automation, ICRA*, 769–774, IEEE. 13, 27, 43

Greene, M.R. & Oliva, A. (2009). The briefest of glances: the time course of natural scene understanding. *Psychological Science*, **20**, 464–472. 98

Greene, N.T., Paige, G.D. & Localization, A.S. (2012). Influence of sound source width on human sound localization. **1**, 6455–6458. 13

Griffiths, T.D. & Warren, J.D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, **5**, 887–892. 1

Grothe, B. (2000). The evolution of temporal processing in the medial superior olive, an auditory brainstem structure. *Progress in Neurobiology*, **61**, 581–610. 2, 9

Grothe, B. (2003). New roles for synaptic inhibition in sound localization. *Nature Reviews Neuroscience*, **4**, 540–550. 10

Grothe, B., Pecka, M. & McAlpine, D. (2010). Mechanisms of sound localization in mammals. *Physiological reviews*, **90**, 983–1012. 2

Guentchev, K. & Weng, J. (1998). Learning-based three dimensional sound localization using a compact non-coplanar array of microphones. In *AAAI Symposium of Intelligent Environments*, Citeseer. 21, 22

Guinan, J.J., Guinan, S.S. & Norris, B.E. (1972a). Single auditory units in the superior olivary complex I: Responses to sounds and classifications based on physiological properties. *International Journal of Neuroscience*, **4**, 101–120. 9

Guinan, J.J., Norris, B.E. & Guinan, S.S. (1972b). Single Auditory Units in the Superior Olivary Complex: II: Locations of Unit Categories and Tonotopic Organization. *International Journal of Neuroscience*, **4**, 147–166. 9

Guo, Y., Wang, X., Wu, C., Fu, Q., Ma, N. & Brown, G.J. (2016). A robust dual-microphone speech source localization algorithm for reverberant environments. In *Interspeech*, 3354–3358. 37

Hafting, T., Fyhn, M., Molden, S., Moser, M.B. & Moser, E.I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, **436**, 801. 25

Harrison, D.G. & De Kamps, M. (2011). A dynamical model of feature-based attention with strong lateral inhibition to resolve competition among candidate feature locations. In *AISB 2011 Convention-Proceedings of the AISB 2011 Symposium on Architectures for Active Vision*, 43–48, Society for the Study of Artificial Intelligence and Simulation of Behaviour. 97

Hartmann, K., Goldenberg, G., Daumüller, M. & Hermsdörfer, J. (2005). It takes the whole brain to make a cup of coffee: the neuropsychology of naturalistic actions involving technical devices. *Neuropsychologia*, **43**, 625–637. 2

Heckmann, M.H.M., Rodemann, T.R.T., Joublin, F.J.F., Goerick, C.G.C. & Scholling, B.S.B. (2006). Auditory Inspired Binaural Robust Sound Source

Localization in Echoic and Noisy Environments. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (2006)*, 368–373. 26

Heinrich, S. & Wermter, S. (2011a). Towards robust speech recognition for human-robot interaction. In *Proceedings of the IROS2011 Workshop on Cognitive Neuroscience Robotics (CNR)*, 29–34. 39

Heinrich, S. & Wermter, S. (2011b). Towards robust speech recognition for human-robot interaction. In *Proceedings of the IROS2011 Workshop on Cognitive Neuroscience Robotics (CNR)*, 29–34. 83

Hengel, P.W.J.V. & Andringa, T.C. (2007). Verbal aggression detection in complex social environments. *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*, 15–20. 1

Hiatt, L.M., Khemlani, S.S. & Trafton, J.G. (2012). An explanatory reasoning framework for embodied agents. *Biologically Inspired Cognitive Architectures*, **1**, 23–31. 11

Hill, K.T., Bishop, C.W. & Miller, L.M. (2012). Auditory grouping mechanisms reflect a sound's relative position in a sequence. *Frontiers in human neuroscience*, **6**, 158. 3

Hofman, P., Van Riswick, J. & Van Opstal, A. (1998). Relearning sound localization with new ears. *Nature Neuroscience*, **1**, 417–421. 5, 12, 81

Holdsworth, J., Nimmo-smith, I., Patterson, R. & Rice, P. (1988). Implementing a GammaTone Filter Bank * The GammaTone filter in the time domain. 1–5. 49

Horimoto, N., Ogawa, T. & Saito, T. (2012). Basic Analysis of Digital Spike Maps. In A.E. Villa, W. Duch, P. Érdi, F. Masulli & G. Palm, eds., *Artificial Neural Networks and Machine Learning – ICANN 2012*, vol. 7552 of *Lecture Notes in Computer Science*, 161–168, Springer. 12

Hornstein, J., Lopes, M., Santos-Victor, J. & Lacerda, F. (2006). Sound Localization for Humanoid Robots - Building Audio-Motor Maps based on the HRTF. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (2006)*, 1170–1176. 27

Hu, J.S.H.J.S., Liu, W.H.L.W.H., Cheng, C.C.C.C.C. & Yang, C.H.Y.C.H. (2006). Location and Orientation Detection of Mobile Robots Using Sound Field Features under Complex Environments. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (2006)*, 1151–1156. 24

Huang, J., Ohnishi, N. & Sugie, N. (1995). A biomimetic system for localization and separation of multiple sound sources. *Transactions on Instrumentation and Measurement*, **44**, 733–738. 21

Huang, J., Ohnishi, N. & Sugie, N. (1997a). Sound localization in reverberant environment based on the model of the precedence effect. *Transactions on Instrumentation and Measurement*, **46**, 842–846. 21

Huang, J., Supaongprapa, T., Terakura, I., Ohnishi, N. & Sugie, N. (1997b). Mobile robot and sound localization. In *Intelligent Robots and Systems*, vol. 2, 683–689, IEEE. 21

Huang, J., Supaongprapa, T. & Wang, F. (1999). A model-based sound localization system and. *Robotics and Autonomous Systems*, **27**, 199–209. 21

Huo, J. & Murray, A. (2009). The adaptation of visual and auditory integration in the barn owl superior colliculus with Spike Timing Dependent Plasticity. *Neural Networks*, **22**, 913–921. 10

Hurmalainen, A., Mahkonen, K., Gemmeke, J.F. & Virtanen, T. (2011). Exemplar-based Recognition of Speech in Highly Variable Noise. 1–5. 3

Hwang, S., Shin, K. & Park, Y. (2006). Artificial ear for robots. In *Conference on Sensors*, 1460–1463, Ieee. 81

Igarashi, J., Shouno, O., Fukai, T. & Tsujino, H. (2011). Real-time simulation of a spiking neural network model of the basal ganglia circuitry using general purpose computing on graphics processing units. *Neural Networks*, **24**, 950–60. 98

Ince, G., Nakadai, K., Rodemann, T., ichi Imura, J., Nakamura, K. & Nakajima, H. (2011a). Incremental learning for ego noise estimation of a robot. In *International Conference on Intelligent Robots and Systems, IROS*, 131–136, IEEE. 38

Ince, G., Nakadai, K., Rodemann, T., Imura, J.i., Nakamura, K. & Nakajima, H. (2011b). Assessment of single-channel ego noise estimation methods. *2011*

IEEERSJ International Conference on Intelligent Robots and Systems, **23**, 106–111. 38

Irie, R. (1995). *Robust sound localization: An application of an auditory perception system for a humanoid robot*. Ph.D. thesis, Massachusetts Institute of Technology. 20

Irvine, D., Park, V. & McCormick, L. (2001). Mechanisms underlying the sensitivity of neurons in the lateral superior olive to interaural intensity differences. *Journal of Neurophysiology*, **86**, 2647. 10, 93

Jeffress, L.A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, **41**, 35. 11, 30

Jenkins, W.M. & Merzenich, M.M. (1984). Role of cat primary auditory cortex for sound-localization behavior. *Journal of Neurophysiology*, **52**, 819–847. 4

Jiang, Y. & Liu, R. (2014). Binaural deep neural network for robust speech enhancement. In *International Conference on Signal Processing, Communications and Computing, ICSPCC*, IEEE. 3

Joris, P., Smith, P. & Yin, T. (1998). Coincidence detection minireview in the auditory system: 50 years after Jeffress. *Neuron*, **21**, 1235–1238. 9, 60

Julian, K., Mern, J. & Tompa, R. (2017). Uav depth perception from visual, images using a deep convolutional neural network. 97

Kanda, T., Ishiguro, H., Imai, M. & Ono, T. (2004). Development and evaluation of interactive humanoid robots. *Proceedings of the IEEE*, **92**, 1839–1850. 95

Karmarkar, U.R. & Buonomano, D.V. (2007). Timing in the absence of clocks: encoding time in neural network states. *Neuron*, **53**, 427–38. 25

Kayser, C., Petkov, C.I., Lippert, M. & Logothetis, N.K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, **15**, 1943–1947. 4

Kelso, J.a.S., Dumas, G. & Tognoli, E. (2013). Outline of a general theory of behavior and brain coordination. *Neural networks: the official journal of the International Neural Network Society*, **37**, 120–31. 11

Kennedy, H. & Dehay, C. (2012). *Self-organization and interareal networks in the primate cortex.*, vol. 195. Elsevier B.V., 1st edn. 93

Keyrouz, F. & Saleh, a.A. (2007). Intelligent Sound Source Localization Based on Head-Related Transfer Functions. *2007 IEEE International Conference on Intelligent Computer Communication and Processing*, 97–104. 27

Kim, D.K.D. (2006). Neural network mechanism for the orientation behavior of sand scorpions towards prey. *IEEE Transactions on Neural Networks*, **17**, 1070–1076. 1

Kim, H.D.K.H.D., Komatani, K. & Ogata, T. (2007). Auditory and Visual Integration based Localization and Tracking of Multiple Moving Sounds in Daily-life Environments. *ROMAN 2007 The 16th IEEE International Symposium on Robot and Human Interactive Communication*, **2**, 399–404. 97

Kim, H.S. & Choi, J.S. (2009). Sound Source Localization Using Sparse Coding and SOM Jong-suk Choi. *Science And Technology*, 1–7. 98

Kim, S.W., Lee, J.Y., Kim, D., You, B.J. & Doh, N.L. (2011). Human localization based on the fusion of vision and sound system. *2011 8th International Conference on Ubiquitous Robots and Ambient Intelligence URAI*, 495–498. 97

Kim, U.H., Nakadai, K. & Okuno, H.G. (2015). Improved sound source localization in horizontal plane for binaural robot audition. *Applied Intelligence*, **42**, 63–74. 55

King, B.Y.A.J. & Palmer, A.R. (1983). 361 cells responsive to free-field auditory stimuli. 361–381. 4

Kitani, E.C., Del-moral hernandez, E. & Silva, L.A. (2012). SOMM – Self-Organized Manifold Mapping. 355–362. 98

Klein, D.J., Depireux, D.A., Simon, J.Z. & Shamma, S.A. (2000). Robust Spectrotemporal Reverse Correlation for the Auditory System: Optimizing Stimulus Design. 85–111. 43, 78

Koch, C. (1993). Computational approaches to cognition: The bottom-up view. *Current opinion in neurobiology*, **3**, 203–208. 11, 12, 13, 94, 95

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics*, **43**, 59–69. 64, 66

Kohonen, T. (1995). Learning vector quantization. In *Self-Organizing Maps*, 175–189, Springer. 64

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, **37**, 52–65. 64

Kourtzi, Z. & Connor, C.E. (2011). Neural representations for object perception: structure, category, and adaptive coding. *Annual Review of Neuroscience*, **34**, 45–67. 2

Krichmar, J.L. (2012). Design principles for biologically inspired cognitive robotics. *Biologically Inspired Cognitive Architectures*, **1**, 73–81. 6, 11

Kumon, M. & Noda, Y. (2011). Active soft pinnae for robots. *2011 IEEERSJ International Conference on Intelligent Robots and Systems*, 112–117. 91

Lakatos, P., Musacchia, G., O'Connel, M.N., Falchier, A.Y., Javitt, D.C. & Schroeder, C.E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron*, **77**, 750–761. 92

Lee, B.g.L.B.g. & Choi, J.C.J. (2010). Multi-source sound localization using the competitive k-means clustering. *Emerging Technologies and Factory Automation ETFA 2010 IEEE Conference on*, **2**. 64

Lee, M., Choi, J. & Park, M. (2009). Design of the Robotic System for Human-Robot Interaction using Sound Source Localization , Mapping Data and Voice Re ecognition JaeM. 1143–1147. 87

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, vol. 10, 707–710. 84

Li, D.L.D. & Levinson, S.E. (2002). A linear phase unwrapping method for binaural sound source localization on a robot. *Proceedings 2002 IEEE International Conference on Robotics and Automation Cat No02CH37292*, **1**, 19–23. 47

Li, Z., Member, S., Herfet, T., Member, S., Grochulla, M. & Thorm, T. (2012). Multiple Active Speaker Localization based on Audio-visual Fusion in two Stages *. 55

REFERENCES

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*. 97

Lim, Y.S.L.Y.S., Choi, J.S.C.J.S. & Kim, M.K.M. (2007). Particle Filter Algorithm for Single Speaker Tracking with Audio-Video Data Fusion. *ROMAN 2007 The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 363–367. 39

Liu, H. & Shen, M. (2010a). Continuous sound source localization based on microphone array for mobile robots. In *International Conference on Intelligent Robots and Systems, IROS*, 4332–4339, IEEE. 28

Liu, H. & Shen, M. (2010b). Continuous sound source localization based on microphone array for mobile robots. In *International Conference on Intelligent Robots and Systems, IROS*, 4332–4339. 85

Liu, J. & Yang, G.Z. (2014). Robust speech recognition in reverberant environments by using an optimal synthetic room impulse response model. *Speech Communication*, **67**, 65–77. 97

Liu, J., Erwin, H., Wermter, S. & Elsaid, M. (2008). A biologically inspired spiking neural network for sound localisation by the inferior colliculus. *International Conference Conference on Artificial Neural Networks, ICANN*, 396–405. 35

Liu, J., Perez-Gonzalez, D., Rees, A., Erwin, H. & Wermter, S. (2009). Multiple sound source localisation in reverberant environments inspired by the auditory midbrain. *International Conference on Artificial Neural Networks, ICANN*, 208–217. 35

Liu, J., Perez-Gonzalez, D., Rees, A., Erwin, H. & Wermter, S. (2010). A biologically inspired spiking neural network model of the auditory midbrain for sound source localisation. *Neurocomputing*, **74**, 129–139. 27, 32, 41, 43, 47, 48, 54, 60, 62, 63, 72, 74

Liu, J., Johns, E. & Yang, G.Z. (2011). A scene-associated training method for mobile robot speech recognition in multisource reverberated environments. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, 542–549, IEEE. 94

Liu, M.Y., Breuel, T. & Kautz, J. (2017). Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, 700–708. 97

Liu, P.R. & Meng, M.Q.H. (2007). A bioinspired robotic sound localization method. *2007 Ieeeasme International Conference on Advanced Intelligent Mechatronics*, 1–6. 20

Lloyd, S. (1982). Least squares quantization in PCM. *Transactions on Information Theory*, **28**, 129–137. 64, 65

Lopez-Poveda, E.A., Palmer, A.R. & Meddis, R. (2010). *The neurophysiological bases of auditory perception*. Springer. 2

Lund, H., Webb, B. & Hallam, J. (1998). Physical and temporal scaling considerations in a robot model of cricket calling song preference. *Artificial Life*, **4**, 95–107. 5

Lv, X. & Zhang, M. (2008). A Sound Source Tracking System Based on Robot Hearing and Vision. *Computer Science and Software*, 1119–1122. 76

Lyon, R. (1983). A computational model of binaural localization and separation. *ICASSP 83 IEEE International Conference on Acoustics Speech and Signal Processing*, **8**, 1148–1151. 19

Ma, W.J., Beck, J.M., Latham, P.E. & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, **9**, 1432–1438. 27

Maas, R., Schwarz, A., Zheng, Y., Reindl, K., Meier, S., Sehr, A. & Kellermann, W. (2011). A Two-Channel Acoustic Front-End for Robust Automatic Speech Recognition in Noisy and Reverberant Environments. 41–46. 37

Maass, W. (1997). Networks of spiking neurons: the third generation of neural network models. *Neural networks*, **10**, 1659–1671. 25

Maass, W. & Bishop, C.M. (2001). *Pulsed neural networks*. MIT press. 25

Maass, W., Natschl, T. & Markram, H. (2002). Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations. **2560**, 2531–2560. 25

REFERENCES

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. of the Berkeley symposium on mathematical statistics and probability*, vol. 1, 14, California, USA. 64

Maeder, P.P., Meuli, R.a., Adriani, M., Bellmann, a., Fornari, E., Thiran, J.P., Pittet, a. & Clarke, S. (2001). Distinct pathways involved in sound recognition and localization: a human fMRI study. *NeuroImage*, **14**, 802–816. 25

Marti, A., Cobos, M. & Lopez, J.J. (2012). Automatic speech recognition in cocktail-party situations: A specific training for separated speech. *The Journal of the Acoustical Society of America*, **131**, 1529–1535. 3

Martinson, E. & Schultz, A. (2007). Robotic Discovery of the Auditory Scene. 10–14. 97

Masterton, R.B. & Imig, T.J. (1984). Neural mechanisms for sound localization. *Annual Review of Physiology*, **46**, 275–287. 4

McCarthy, J. (1960). *Programs with common sense*. RLE and MIT computation center. 1

McCarthy, J. & Hayes, P.J. (1981). Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, 431–450, Elsevier. 1

McNaughton, B.L., Battaglia, F.P., Jensen, O., Moser, E.I. & Moser, M.B. (2006). Path integration and the neural basis of the 'cognitive map'. *Nature Reviews Neuroscience*, **7**, 663–678. 25

Meddis, R., Lopez-Poveda, E., Fay, R.R. & Popper, A.N. (2010). *Computational models of the auditory system*, vol. 35. Springer. 31, 32

Metta, G., Sandini, G., Vernon, D., Natale, L. & 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*, 50–56, ACM. 11, 13, 80, 94

Middlebrooks, J. & Green, D. (1991). Sound localization by human listeners. *Annual review of psychology*, **42**, 135–159. 3, 5, 29, 90

Milford, M. & Wyeth, G. (2009). Persistent Navigation and Mapping using a Biologically Inspired SLAM System. *The International Journal of Robotics Research*, **29**, 1131–1153. 25

Milford, M.J., Wyeth, G.F. & Prasser, D. (2004). RatSLAM: a hippocampal model for simultaneous localization and mapping. *IEEE International Conference on Robotics and Automation 2004 Proceedings ICRA 04 2004*, **1**, 403–408. 25

Minato, T., Shimada, M., Ishiguro, H. & Itakura, S. (2004). Development of an android robot for studying human-robot interaction. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 424–434, Springer. 87

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, **49**, 8–30. 1

Møller, M.F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, **6**, 525–533. 67, 68

Moore, B.C. (2012). *An introduction to the psychology of hearing*. Brill. 2

Mori, M. (1970). The uncanny valley. *Energy*, **7**, 33–35. 87

Murray, J., Erwin, H. & Wermter, S. (2009). Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. *Neural Networks*, **22**, 173–189. 27

Murray, J.C., Erwin, H.R. & Wermter, S. (2004). Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. *AI Workshop on NeuroBotics*, **22**, 173–189. 27

Nakadai, K., Lourens, T., Okuno, H.G. & Kitano, H. (2000). Active audition for humanoid. In *AAAI/IAAI*, 832–839. 20

Nakadai, K., Takahashi, T., Okuno, H.G., Nakajima, H., Hasegawa, Y. & Tsujino, H. (2010). Design and implementation of robot audition system'hark'—open source software for listening to three simultaneous speakers. *Advanced Robotics*, **24**, 739–761. 20

REFERENCES

Nakamura, K., Nakadai, K., Asano, F. & Ince, G. (2011). Intelligent sound source localization and its application to multimodal human tracking. *2011 IEEERSJ International Conference on Intelligent Robots and Systems*, **559**, 143–148. 20

Nakashima, H. & Mukai, T. (2005). 3D sound source localization system based on learning of binaural hearing. In *International Conference on Systems, Man and Cybernetics*, vol. 4, 3534–3539, IEEE. 2

Nakashima, H., Mukai, T. & Ohnishi, N. (2002). Self-organization of a sound source localization robot by perceptual cycle. *Proceedings of the 9th International Conference on Neural Information Processing 2002 ICONIP 02*, **2**, 834–838. 92, 98

Natale, L., Metta, G. & Sandini, G. (2002). Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head. *Robotics and Autonomous Systems*, **39**, 87–106. 76

Newell, A., Simon, H.A. *et al.* (1972). *Human problem solving*, vol. 104. Prentice-Hall Englewood Cliffs, NJ. 1

Newton, M.J. & Smith, L.S. (2011). Biologically-inspired neural coding of sound onset for a musical sound classification task. *The 2011 International Joint Conference on Neural Networks*, 1386–1393. 21

Nguyen, Q., Girin, L., Bailly, G., Elisei, F. & Nguyen, D.C. (2018). Autonomous Sensorimotor Learning for Sound Source Localization by a Humanoid Robot. In *Workshop on Crossmodal Learning for Intelligent Robotics in conjunction with IEEE/RSJ IROS*, Madrid, Spain. 13

Nix, J. & Hohmann, V. (2006). Sound source localization in real sound fields based on empirical statistics of interaural parameters. *The Journal of the Acoustical Society of America*, **119**, 463. 26, 54

Nodal, F.R., Kacelnik, O., Bajo, V.M., Bizley, J.K., Moore, D.R. & King, A.J. (2010). Lesions of the Auditory Cortex Impair Azimuthal Sound Localization and Its Recalibration in Ferrets. *Journal of Neurophysiology*, **103**, 1209–1225. 13, 97

Noë, A. & Regan, J.K.O. (2000). Perception , Attention and the Grand Illusion. *Philosophy*, **6**, 6–15. 13

Nunes, L.O., Martins, W.A., Lima, M.V., Biscainho, L.W., Costa, M.V., Goncalves, F.M., Said, A. & Lee, B. (2014). A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays. *Signal Processing, IEEE Transactions on*, **62**, 5171–5183. 25

Okuno, H.G., Ogata, T. & Komatani, K. (2007). Computational Auditory Scene Analysis and Its Application to Robot Audition: Five Years Experience. *Second International Conference on Informatics Research for Development of Knowledge Society Infrastructure ICKS07*. 94

Oliver, D.L., Beckius, G.E., Bishop, D.C., Loftus, W.C. & Batra, R. (2003). Topography of interaural temporal disparity coding in projections of medial superior olive to inferior colliculus. *Journal of Neuroscience*, **23**, 7438–7449. 9, 28

Panchev, C. & Wermter, S. (2006). Temporal sequence detection with spiking neurons: towards recognizing robot language instructions. *Connection Science*, **18**, 1–22. 4

Park, J. & Sandberg, I.W. (1991). Universal approximation using radial-basis-function networks. *Neural computation*, **3**, 246–257. 67

Park, T.J., Klug, A., Holinstat, M. & Grothe, B. (2004). Interaural level difference processing in the lateral superior olive and the inferior colliculus. *Journal of Neurophysiology*, **92**, 289–301. 9

Pavlidi, D., Griffin, A., Puigt, M. & Mouchtaris, A. (2013). Real-time multiple sound source localization and counting using a circular microphone array. *Audio, Speech, and Language Processing, IEEE Transactions on*, **21**, 2193–2206. 28

Perisa, D., Ivancevic, B. & Jambrosic, K. (2004). Sound localization. *Proceedings Elmar2004 46th International Symposium on Electronics in Marine*, **11**, 683–689. 79

Pfeifer, R., Lungarella, M. & Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. *science*, **318**, 1088–1093. 6

Phillips, M.A., Colonnese, M.T., Goldberg, J., Lewis, L.D., Brown, E.N. & Constantine-Paton, M. (2011). A synaptic strategy for consolidation of convergent visuotopic maps. *Neuron*, **71**, 710–724. 97

Pujol, R., Blatrix, S., Le Merre, S., Pujol, T., Chaix, B., Rubel, E.W., Gil-Loyzaga, P., Trigueiros Cunha, N., Puel, J.L., Lenoir, M., Camilleri, M., Daruty de Grandpre, V., Lorenzi, A., Vessigaud, M.A., Lazard, D., Venail, F., Marozeau, J., Stone, J., Lewis, R., Irving, S., Faulconbridge, R., Ewbank, J., Owens, K. & Ewbank, T. (2019). Journey into the world of hearing. `http://www.cochlea.org`. 5

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? in search of meaning in modules and neuron circuits. *Brain and language*, **127**, 86–103. 6, 11

Rao, R.P.N. (2004). Bayesian computation in recurrent neural circuits. *Neural Computation*, **16**, 1–38. 32

Rascon, C. & Meza, I. (2017). Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, **96**, 184–210. 19

Recanzone, G. & Sutter, M. (2008). The biological basis of audition. *Annual Review of Psychology*, **59**, 119–142. 10

Ren, M. & Zou, Y.X. (2012). A novel multiple sparse source localization using triangular pyramid microphone array. *Signal Processing Letters, IEEE*, **19**, 83–86. 28

Richter, C.P., Evans, B.N., Edge, R. & Dallos, P. (1998). Basilar membrane vibration in the gerbil hemicochlea. *Journal of Neurophysiology*, **79**, 2255–2264. 6, 62

Roberts, M.T. & Golding, N.L. (2012). Gabab receptors sharpen tuning of a sound localization circuit. *The Journal of physiology*, **590**, 2951–2952. 9

Rodemann, T. (2010). A study on distance estimation in binaural sound localization. *Intelligent Robots and Systems IROS 2010 IEEERSJ International Conference on*, 425–430. 26

Rodemann, T., Heckmann, M., Joublin, F., Goerick, C. & Scholling, B. (2006). Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. In *International Conference on Intelligent Robots and Systems, IROS*, 860–865, IEEE. 26

Rokni, U. & Sompolinsky, H. (2012). How the brain generates movement. *Neural Computation*, **24**, 289–331. 26, 59, 95

Roman, N., Wang, D. & Brown, G. (2003). Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, **114**, 2236–2252. 2, 38, 54

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, **65**, 386. 67

Rouat, J., Loiselle, S. & Molotchnikoff, S. (2011). Variable frame rate hierarchical analysis for robust speech recognition. *2011 IEEERSJ International Conference on Intelligent Robots and Systems*, **40**, 518–523. 37

Rubruck, A., Aminian, A., Yalpi, J., Hanzal, P., Winde, S., Kuppusami, S., Younis, S., Thomas, S., Strahl, E., Bauer, J., Davila-Chacon, J., Heinrich, S. & Wermter, S. (2013a). CoCoCo, Coffee Collecting Companion. In *Proceedings of the 28th Conference on Artificial Intelligence, AAAI*, In press, AAAI Press. 84

Rubruck, A., Aminian, A., Yalpi, J., Hanzal, P., Winde, S., Kuppusami, S., Younis, S., Thomas, S., Strahl, E., Bauer, J., Davila-Chacon, J., Heinrich, S. & Wermter, S. (2013b). CoCoCo, Coffee Collecting Companion. In *Proceedings of the 28th Conference on Artificial Intelligence, AAAI*, In press. 94

Rucci, M., Tononi, G. & Edelman, G.M. (1997). Registration of Neural Maps through Value-Dependent Learning: Modeling the Alignment of Auditory and Visual Maps in the Barn Owl's Optic Tectum. **17**, 334–352. 57

Rucci, M., Wray, J. & Edelman, G.M. (2000). Robust localization of auditory and visual targets in a robotic barn owl. **30**, 181–193. 57

Ruesch, J., Lopes, M., Bernardino, a., Hornstein, J., Santos-Victor, J. & Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub. *2008 IEEE International Conference on Robotics and Automation*, 962–967. 92

Ruggles, D., Bharadwaj, H. & Shinn-Cunningham, B.G. (2011). Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 15516–15521. 2, 92

Russell, S. & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edn. 93

Sagi, B., Nemat-Nasser, S.C., Kerr, R., Hayek, R., Downing, C. & Hecht-Nielsen, R. (2001). A biologically motivated solution to the cocktail party problem. *Neural Computation*, **13**, 1575–1602. 37

Salb, S. & Duhr, P. (2009). Vergleich der Soundman OKM II Studio Klassik mit dem Neumann-Kunstkopf KU 81i nach technischen und klanglichen aspekten. Tech. rep., SAE Institute and University of Middlesex. 13

Salminen, N.H., Tiitinen, H., Miettinen, I., Alku, P. & May, P.J.C. (2010). *Asymmetrical representation of auditory space in human cortex*, vol. 1306. Elsevier B.V. 10

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

Sasaki, Y., Kabasawa, M., Thompson, S., Kagami, S. & Oro, K. (2012). Spherical Microphone Array for Spatial Sound Localization for a Mobile Robot. 713–718. 24

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., Kamvar, M. & Strope, B. (2010). "your word is my command": Google search by voice: A case study. In *Advances in Speech Recognition*, 61–90, Springer. 83

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, **49**, 336–347. 13

Schauer, C. & Gross, H. (2003). A Computational Model of Early Auditory – Visual Integration. 362–369. 14

Schenkman, B.N. & Nilsson, M.E. (2011). Human echolocation: Pitch versus loudness information. *Perception*, **40**, 840–852. 2

Schierwagen, A. (2012). On reverse engineering in the cognitive and brain sciences. *Natural Computing*, **11**, 141–150. 11

Schnupp, J., Nelken, I. & King, A. (2011). *Auditory neuroscience: Making sense of sound*. The MIT Press. 2, 5, 7, 9, 10, 31, 32, 44

Shannon, C.E. (1948). The mathematical theory of communication. 1963. *MD computing computers in medical practice*, **14**, 306–17. 12

Shinn-Cunningham, B.G. (2008). Object-based auditory and visual attention. *Trends in cognitive sciences*, **12**, 182–186. 11

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D. & Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *ICML*. 97

Sinapov, J., Bergquist, T., Schenck, C., Ohiri, U., Griffith, S. & Stoytchev, a. (2011). Interactive object recognition using proprioceptive and auditory feedback. *The International Journal of Robotics Research*, **30**, 1250–1262. 1, 97

Singheiser, M., Gutfreund, Y. & Wagner, H. (2012). The representation of sound localization cues in the barn owl's inferior colliculus. *Frontiers in neural circuits*, **6**, 45. 13

Siracusa, M., Morency, L.P., Wilson, K., Fisher, J. & Darrell, T. (2003). A multi-modal approach for determining speaker location and focus. In *Proceedings of the 5th international conference on Multimodal interfaces*, 77–80, ACM. 20

Slaney, M. (1993). An efficient implementation of the Patterson-Holdsworth auditory filter bank. Tech. rep., Apple Computer, Perception Group. 26, 29, 49, 59, 61, 75

Smith, P., Joris, P. & Yin, T. (1993). Projections of physiologically characterized spherical bushy cell axons from the cochlear nucleus of the cat: Evidence for delay lines to the medial superior olive. *The Journal of Comparative Neurology*, **331**, 245–260. 9

Spille, C., Dietz, M., Hohmann, V. & Meyer, B.T. (2013). Binaural scene analysis and automatic speech recognition. 3

Stein, B. & Meredith, M. (1993a). *The merging of the senses*. The MIT Press. 2

Stein, B.E. & Meredith, M.A. (1993b). *The Merging Of The Senses*. Cognitive Neuroscience Series, MIT Press, 1st edn. 10

Stein, R.B. (1967). Some models of neuronal variability. *Biophysical journal*, **7**, 37–68. 10, 41

REFERENCES

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A. & Goodrich, M. (2006). Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 33–40, ACM. 88

Stork, H.G. (2012). Towards a scientific foundation for engineering cognitive systems–a european research agenda, its rationale and perspectives. *Biologically Inspired Cognitive Architectures*, **1**, 82–91. 11

Stramadinoli, F., Ruciński, M., Znajdek, J., Rohlfing, K.J. & Cangelosi, A. (2011). From sensorimotor knowledge to abstract symbolic representations. *Procedia Computer Science*, **7**, 269–271. 2

Sutton, R.S. & Barto, A.G. (1998). *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge. 97

Tamai, Y., Sasaki, Y., Kagami, S. & Mizoguchi, H. (2005). Three ring microphone array for 3D sound localization and separation for mobile robot audition. In *International Conference on Intelligent Robots and Systems, IROS*, 4172–4177, IEEE. 24

Teramoto, W., Sakamoto, S., Furune, F., Gyoba, J. & Suzuki, Y. (2012). Compression of auditory space during forward self-motion. *PloS one*, **7**, e39402. 97

Thompson, E.R. & Dau, T. (2008). Binaural processing of modulated interaural level differences. *Journal of the Acoustical Society of America*, **123**, 1017–1029. 10

Titze, I.R. & Martin, D.W. (1998). Principles of voice production. 49

Trifa, V.M., Koene, a., Moren, J. & Cheng, G. (2007). Real-time acoustic source localization in noisy environments for human-robot multimodal interaction. *ROMAN 2007 The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 393–398. 94

Twiefel, J., Baumann, T., Heinrich, S. & Wermter, S. (2014). Improving domain-independent cloud-based speech recognition with domain-dependent phonetic post-processing. In *AAAI*, In press. 39, 82, 84

Valin, J., Michaud, F., Rouat, J. & Létourneau, D. (2003). Robust sound source localization using a microphone array on a mobile robot. In *International Conference on Intelligent Robots and Systems, IROS*, vol. 2, 1228–1233, IEEE. 23, 24

van der Zant, T. & Iocchi, L. (2011). Robocup@ home: Adaptive benchmarking of robot bodies and minds. *Social Robotics*, 214–225. 2

van Hateren, J.H. (1992). A theory of maximizing sensory information. *Biological cybernetics*, **68**, 23–29. 13

Van Rijsbergen, C. (1979). Evaluation. In *Information Retrieval*, Butterworths, London. 68

Vasilkov, V. & Tikidji-Hamburyan, R. (2012). Accurate Detection of Interaural Time Differences by a Population of Slowly Integrating Neurons. *Physical Review Letters*, **108**, 1–5. 10

Voutsas, K. & Adamy, J. (2007). A biologically inspired spiking neural network for sound source lateralization. *Transactions on Neural Networks*, **18**, 1785–1799. 25, 26, 54

Wagner, K. & Dobkins, K.R. (2011). Synaesthetic associations decrease during infancy. *Psychological Science*, **22**, 1067–1072. 97

Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P. & Woelfel, J. (2004). Sphinx-4: A flexible open source framework for speech recognition. 83

Wang, Z.Q., Zhang, X. & Wang, D. (2018). Robust tdoa estimation based on time-frequency masking and deep neural networks. In *Proc. Interspeech*, 322–326. 38

Weintraub, M. (1986). A computational model for separating two simultaneous talkers. *ICASSP 86 IEEE International Conference on Acoustics Speech and Signal Processing*, **11**, 81–84. 37

Wermter, S., Palm, G., Weber, C. & Elshaw, M. (2005). Towards biomimetic neural learning for intelligent robots. In *Biomimetic Neural Learning for Intelligent Robots*, 1–18, Springer. 11

West, L.J., Epstein, W. & Dember, W.N. (2018). Perception. 2

Willert, V., Eggert, J., Adamy, J., Stahl, R. & Körner, E. (2006). A probabilistic model for binaural sound localization. *Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, **36**, 982–994. 26, 27, 44, 54

REFERENCES

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic bulletin & review*, **9**, 625–636. 11, 13

Winston, P.H. (2012). The next 50years: A personal view. *Biologically Inspired Cognitive Architectures*, **1**, 92–99. 11

Woodruff, J. & Wang, D. (2013). Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues. *Audio, Speech, and Language Processing, IEEE Transactions on*, **21**, 806–815. 39

Wright, B.a. & Zhang, Y. (2006). A review of learning with normal and altered sound-localization cues in human adults. *International Journal of Audiology*, **45 Suppl 1**, S92–S98. 1

Yamamoto, S., Nakadai, K., Tsujino, H. & Okuno, H.G. (2004). Assessment of general applicability of robot audition system by recognizing three simultaneous speeches. *2004 IEEERSJ International Conference on Intelligent Robots and Systems IROS IEEE Cat No04CH37566*, **3**, 2111–2116. 3

Zhang, X. & Wang, D. (2017). Deep learning based binaural speech separation in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25**, 1075–1084. 37

Zhao, Y., Wang, D., Johnson, E.M. & Healy, E.W. (2018). A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions. *The Journal of the Acoustical Society of America*, **144**, 1627–1637. 13, 95

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z. *et al.* (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, **77**, 980–991. 92

Declaration on Oath

Eidesstattliche Versicherung

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, 9^{th} of March, 2019 — Jorge Dávila Chacón

City and Date
Ort und Datum

Signature
Unterschrift