ANNÉE 2016

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Traitement du signal et telecommunications
Ecole doctorale MATISSE

présentée par

Aly Magassouba

préparée à l'unité de recherche IRISA – UMR6074
Institut de Recherche en Informatique et Système Aléatoires
ISTIC

Aural servo:
towards an alternative approach to sound localization for robot motion control

Thèse soutenue à Rennes le 5 décembre 2016

devant le jury composé de :

Patrick DANES
Professeur à l'Université Paul Sabatier de Toulouse / *Rapporteur*

Emmanuel Vincent
Directeur de recherche Inria, Nancy / *Rapporteur*

Ivan PETROVIC
Professeur à l'Université de Zagreb, Croatie / *Examinateur*

Radu HORAUD
Directeur de recherche Inria, Grenoble / *Examinateur*

Bruno GAS
Professeur à l'Université Pierre et Marie Curie, Paris / *Examinateur*

Nancy BERTIN
Chargée de recherche CNRS, Rennes / *Co-directrice de thèse*

François CHAUMETTE
Directeur de recherche Inria, Rennes / *Directeur de thèse*

Synthèse

Est-ce que les robots doivent nécessairement localiser les sources sonores afin d'interagir avec leur environement ?
Cette question a été le fil central de cette thèse. En effet, dans un contexte d'interactions homme-robot, un robot doté d'un sens de l'ouïe fiable est essentiel. L'information auditive en robotique est généralement exploitée par des capacités tels que la localisation de source, la séparation de sources, le traitement de la parole et l'analyse de scène acoustique, qui définissent le sujet de *l'audition robotique*. Dans cette thématique, une attention particulière a été accordée à la localisation de source, principalement parce qu'elle est la première étape de l'interaction auditive.

Les techniques de localisation de source, en robotique, s'inspirent des capacités auditives des mammifères et plus particulièrement de celles des humains. L'approche typique de la localisation de source en robotique consiste à transposer les connaissances acquises en psychoacoustique et en physiologie dans un système d'audition artificiel dont sont extraits des indices sonores. Ces méthodes supposent généralement que les capteurs et les sources d'intérêt sont statiques. Les indices sonores tels que la différence interaurale en temps et la différence interaurale en amplitude ou les densités spectrales, fournissent des informations sur la direction d'arrivée du son dans le plan horizontal et le plan vertical. L'efficacité du processus de localisation dépend uniquement de la précision des indices sonores extraits.

Cependant, imiter le système auditif humain se trouve être relativement complexe en environnement réel. D'une part, la perception et plus particulièrement les indices sonores sont influencés par la forme du corps (tête, pavillon auditif, torse ...) et d'autre part par les conditions acoustiques (acoustiques de la salle, bruit, réverbération). Ces indices sonores varient d'un individu à un autre et d'un environement à un autre. Par conséquent, même les systèmes artificiels les plus aboutis ne peuvent être déployés que dans des environnements contrôlés. En environnement réel, qui implique la présence de réverbération, de bruit et de mouvement, la précision des indices sonores se dégrade considérablement, ce qui entraine des erreurs de localisation.

C'est la raison pour laquelle, une nouvelle ligne d'approches qui étudie en profondeur les capacités du système auditif humain émerge : se baser uniquement sur la pertinence et la précision des indices sonores ne suffit pas pour une localisation robuste. Plusieurs analyses des capacités de localisation chez l'humain, telles que celles proposées dans [Pop08] ou [CLH97], montrent une grande variabilité de précision lorsqu'il s'agit d'estimer la direction d'arrivée du son. Dans ces expériences,

le processus de localisation démontre une bonne précision pour les stimuli sonores situés en face du sujet alors que les positions plus latérales entraînent des erreurs de localisation plus importantes. D'autres études [ALO90, CTS68] ont rapporté que l'estimation de la distance a tendance à être largement sous-estimé. Néanmoins, malgré ces imprécisions, des actions telles que tourner la tête en direction de la source sonore sont exécutées avec précision et sans effort particulier. Ces résultats tendent à démontrer que la précision des indices sonores n'est pas cruciale pour les interactions basées sur l'ouïe. Cette hypothèse est également corroborée par une étude de [HCM83], qui a montré que des sujets souffrant de perte auditive ont développé des stratégies de localisation basée sur des indices sonores erronés, en utilisant le mouvement de leur tête et leur sens visuel. Leur système auditif s'est adapté à l'utilisation d'indices sonores incorrects.

A partir de ces résultats, on peut alors suggérer que les points-clés du système auditif humain ne sont pas la précision des indices sonores, mais plutôt la capacité d'adaptation et la flexibilité par rapport aux conditions acoustiques comme le montre [HVRVO98]. Un système auditif efficace peut être défini comme un mécanisme qui combinent différentes modalités telles que la vision (via des références spatiales), le mouvement et la mémoire des expériences auditives précédentes. L'espace acoustique est ensuite appréhendé dynamiquement et les erreurs potentielles des indices sonores peuvent être compensées par ces modalités. Différentes approches peuvent alors être considérées comme la localisation combinant l'ouïe et la vision, la localisation par apprentissage, ou par le biais de mouvement.

Cette thèse se concentre plus particulièrement sur les stratégies de mouvement liées à la perception auditive, dans un contexte binaural. Ces stratégies ont été explorées par les méthodes d'*audition active*. L'*audition active* consiste à mesurer les indices sonores en générant un mouvement prédéfini du robot afin d'améliorer leur précision. L'étape de localisation est effectuée ensuite par la fusion des mesures prises à partir des différentes positions du robot. Cette approche améliore le processus de localisation, mais fait face à certaines limites. D'abord, les mouvements prédéfinis peuvent limiter ou modifier l'interaction avec un comportement non-naturel du robot. Ensuite, les scènes dynamiques, où le robot et/ou la source sont en mouvement nécéssitent des modélisations supplémentaires (généralement la dynamique de la source) et un tracking de sorte que la fusion des mesures reste cohérente. Enfin, plus important encore, l'audition active repose toujours sur la localisation et la précision des indices sonores. Sachant que le robot lui-même et les conditions acoustiques influent sur les indices sonores à chaque position du robot dans la scène, cette approche est limitée par une bonne modélisation des conditions acoustiques, de façon similaire aux techniques de localisation en configuration statique.

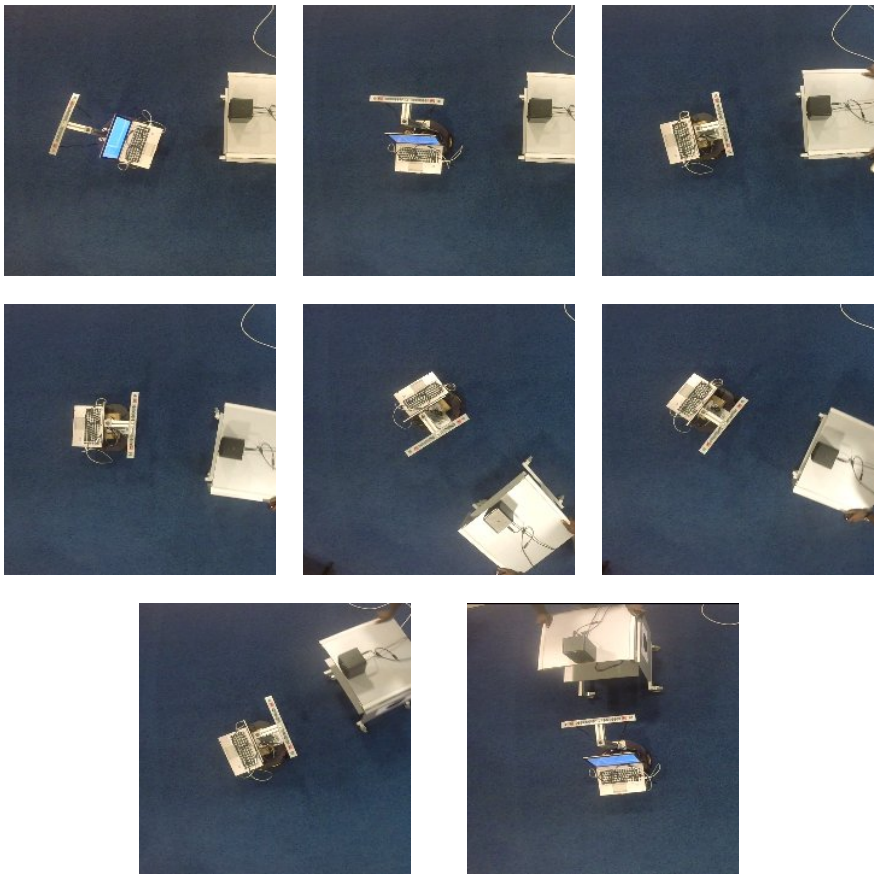
En réponse à la question initiale sur la nécessité de la localisation, le travail développé dans cette thèse diffère des approches cités auparavant en définissant une interaction auditive, comme une tâche modélisée par une commande référencée capteurs. Cette méthode que nous appelons asservissement sonore, ne localise pas la source sonore. A la place, on définit une tâche par des conditions de mesures à satisfaire, qui correspondent généralement à une position particulière du robot. Par exemple, une tâche en asservissement sonore, consiste à positionner le robot afin

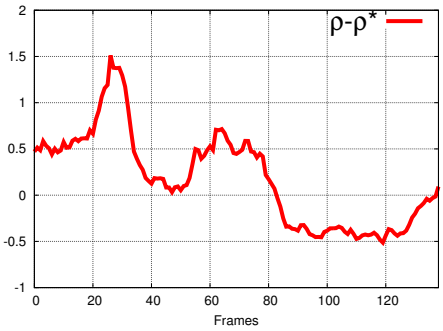
de satisfaire des conditions définies par un ensemble de mesures auditives, alors qu'une approche basée sur la localisation nécessite d'extraire les angles d'azimut et d'altitude et la distance par rapport à la source sonore, avant de déplacer le robot à l'emplacement désiré. Dans l'asservissement sonore, le mouvement du robot est généré en temps réel, à travers une boucle de commande basée sur des indices sonores mesurés à la volée. La commande du robot est obtenue en modélisant la relation entre la dynamique des indices sonores et le mouvement des capteurs auditifs. De cette façon, notre approche repose principalement sur la dynamique des indices sonores et non sur leur valeur intrinsèque, ce qui permet une meilleure tolérance vis-à-vis des conditions acoustiques. Avoir des indices sonores précis n'est alors plus requis. En conséquence, l'étape de localisation est ignorée et la modélisation des conditions acoustiques ou des perturbations de la tête ne sont pas nécessaires. De la même facon, il est possible d'effectuer des tâches basées sur des indices sonores erronés du moment que la dynamique de ces indices reste cohérente. Cette approche appliquée au domaine de l'audition robotique est la contribution essentielle de cette thèse. Le manuscrit est donc centré sur la description de cette approche et comprend la description de la commande référencée capteurs, la modélisation des indices sonores pour la commande et des résultats expérimentaux qui valident la pertinence et l'utilité de cette méthode en environnement réel. Ces expériences sont réalisées dans des configurations dynamiques en considérant des micros en champ libre et des micros intégrés dans des robots humanoïdes.

Dans cette thèse, nous avons introduit plusieurs lois de commandes se basant sur la variation des indices sonores par rapport aux mouvements du robot :

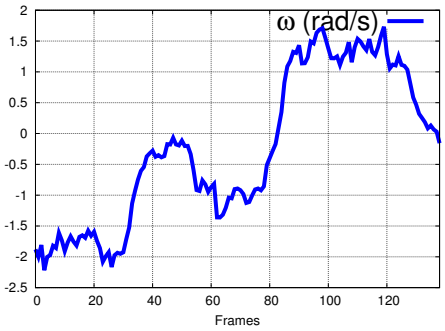
- une première basée sur la différence interaurale en temps (ITD),
- une deuxième basée sur la différence interaurale en amplitude (ILD),
- ut une dernière utilisant le niveau d'energie du son.

A partir de ces commandes, nous avons expérimentés différentes tâches de positionnement. Que ce soit avec l'ILD ou avec l'ITD, les tâches qui consistent à orienter le robot vis-à-vis d'une source sonore ont été effectuées dans des conditions réelles (de bruit et de réverbération). De plus, les avantages de ces méthodes ont été soulignés par la résolution des problèmes liés à la perception auditive. Dans le cas de l'ILD, l'ambiguïté avant/arrière est intrinsèquement résolue par le système de contrôle. Dans le cas de l'ITD, nous avons démontré la robustesse du système de contrôle que ce soit en champs lointain ou en champs proche. Finalement, nous avons également prouvé que le positionnement en distance en utilisant le niveau d'énergie peut être effectué avec précision dans des conditions réelles. En outre, ces approches sont adaptées aux scènes dynamiques, qui sont rarement abordées en audition robotique. L'experience décrite en Figure 1 résume le potentiel de l'asservissement sonore. Dans celle-ci, le robot est capable de faire face à la source sonore alors que celle-ci est initialement présente dans le dos du robot. Par la suite, lorsque la source est déplacée, le robot s'oriente en temps réel vers celle-ci en la maintenant toujours de face. Cette expérience qui s'est déroulée en environnment réel démontre la robustesse de notre





(i) Erreur

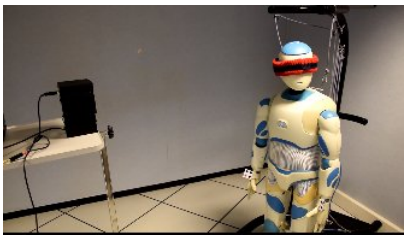


(j) Commande

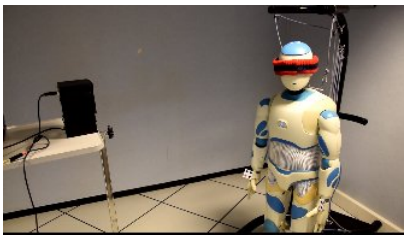
Figure 1: Une expérience illustrant l'asservissement sonore. D'une position quelconque le robot est capable de se retourner vers la source sonore et suivre le mouvement de celle-ci en utilisant deux micros. L'approche utilisée ici est basée sur la différence interaurale en amplitude.

approche. Outre le fait de souligner la pertinence de l'asservissement sonore en conditions réelles, ces expériences valident également la polyvalence de l'asservissement sonore. En effet, ce type d'expérience peut être réalisé, de la même manière, sur

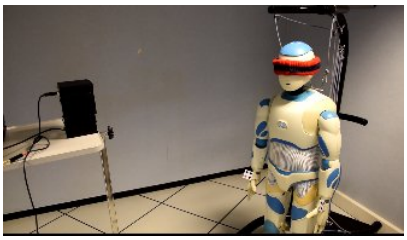
des robots humanoïdes où les microphones sont intégrés à la tête. Cette dernière possibilité est illustrée par la Figure 2 où le robot *Romeo* oriente sa tête vers la source sonore. A notre connaissance, ce travail est l'une des premières tentatives à générer le mouvement d'un robot par rapport à l'information auditive, via une loi de commande explicite.



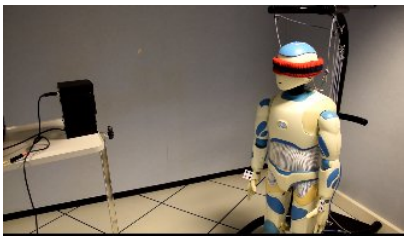
(a)



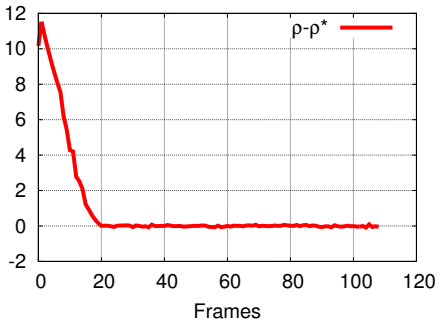
(b)



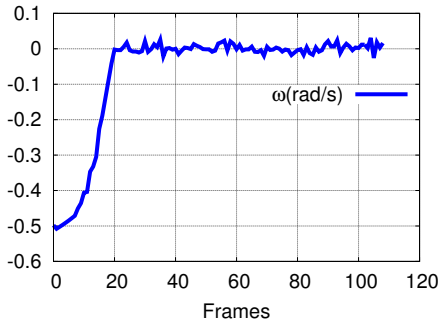
(c)



(d)



(e) Erreur



(f) Commande

Figure 2: *Romeo* oriente sa tête correctement vers la source sonore en utilisant les indices sonores produient par la différence interaurale en temps.

Si l'on se restreint aux systèmes binauraux, les résultats expérimentaux exposés tout au long de cette thèse ne sont pas réalisables par l'état de l'art de l'audition robotique. Au-delà de ce travail, nous aimerions envisager cette approche comme un cadre général qui peut être appliqué à tout type de robots et de conditions acoustiques. A cette fin, les méthodes présentées dans cette thèse pourraient être améliorées afin de gérer des sources sonores non continues ou le suivi et l'évaluation du nombre de sources actives.

Contents

Introduction

Do robots necessarily need to localize sound source(s) in order to engage into interaction? This question builds up the central thread of this thesis. Indeed, towards the dream of autonomous human-robot interactions, a robot endowed with a reliable sense of hearing is essential. Nowadays, this sense of hearing is generally exploited through auditory capabilities such as sound localization, along with sound separation, speech processing and auditory scene analysis, that build up the topic of *robot audition*. In this topic, particular attention has been given to sound source localization, certainly because it is the first step of auditory-based interaction. The localization step allows to control robot motions for positioning tasks such as facing or approaching a speaker in the scene.

Binaural sound source localization techniques are inspired by the way mammals and more particularly humans sense and perceive sound. Typical approaches for sound localization consist in transposing the knowledge in psychoacoustics and physiology into an artificial hearing system that extracts auditory cues. Auditory cues such as the interaural time difference and the interaural level difference or the spectral notches, provide information about the direction of arrival of the sound in the horizontal and the vertical plane. As a result the efficiency of the localization process depends exclusively on the accuracy of the extracted auditory cues.

However mimicking the human hearing system happens to be complex in realistic configurations. Perception and more particularly auditory cues are influenced by the shape of the body (head, pinna, torso...) along with acoustic conditions (acoustics of the room, noise, reverberation). The sensitivity to auditory events is variable for each individual and each location. As a consequence, even the most complex artificial auditory systems can be deployed in controlled environments only. Real world configurations, that include dynamic scenes, reverberation, noises, degrade drastically the accuracy of the auditory cues and at the same time the localization performance.

Still, new lines of approaches have emerged by studying in depth the capabilities of the human hearing system: relying only on the relevance and precision of the auditory cues does not provide enough robustness. Actually a typical analysis of the performance of localization by human listeners as proposed by [Pop08] or [CLH97] shows the variability of accuracy in estimating the direction of sound arrival. In these experiments the localization process showed good accuracy for sound stimuli located around the head front sight while eccentric positions of the source lead to more substantial errors. On top of that, other studies [ALO90, CTS68] reported inaccurate

distance estimations, which tends to be largely underestimated. Nevertheless despite this lack of accuracy, tasks such as head-turn reflex, which consists in turning the head in the direction of the sound source are performed accurately and effortlessly by human listeners. Thus, it could be deduced that accurate auditory cues may not be essential for auditory-based interactions. This assumption is also corroborated by a study of [HCM83], that showed how subjects suffering from hearing loss developed localization strategies based on erroneous auditory cues, by using body motion and visual searching. By contrast, when hearing aids were used in order to recover correct auditory cues, the test subjects localization performance was immediately degraded. Their hearing system was adapted to incorrect auditory cues.

In view of these results, it can be suggested that the key points of human auditory system are not the accuracy of auditory cues but rather the adaptability and flexibility to acoustic configurations as shown by [HVRVO98]. Actually an efficient hearing system can be defined as a mechanism that combines different input modalities, such as additional visual cues (spatial references), motions and memory of previous auditory experiences. The acoustic space is then apprehended dynamically and potential auditory cues errors can be compensated by the latter modalities. Different paradigms can then be considered such as audio-visual localization, localization through learning-based approach, or through efficient motion strategies.

This thesis focuses more particularly on motion strategies governed by auditory perception, in a binaural context. Such strategies have been explored in the so-called *active audition* [NOK00b]. This method consists in measuring the auditory cues actively in order to improve their accuracy, by generating a predefined motion of the robot. The localization step is performed afterwards by fusing the measurements taken from different poses of the robot. This approach improves static sound localization but faces some limitations. Active audition aims to localize sound sources. Knowing that robots and acoustic conditions influence the auditory cues, this approach is limited by a good modelling of the acoustic environment, similarly to sound source localization in a static configuration. Furthermore, dynamic scenes (*i.e.* moving sensors/ sources) need additional processing such as tracking so that auditory cues measurements remains consistent.

As an answer of the initial question about the necessity of localization, the work developed in this thesis differs from the approaches discussed before by defining an auditory-based interaction as a task modelled in a sensor-based framework. This framework, that we term *aural servo* (AS), does not localize sound sources. A task is characterized by a state of measurements to satisfy, that generally corresponds to a particular pose of the robot. For instance, in AS, a homing task consists in positioning the robot in order to satisfy given conditions defined by a set of auditory measurements, while a localization-based approach requires to extract the bearing angles and the distance to the sound source, before moving the robot to the desired location. In AS paradigm, robot motion is generated in real time, through a control loop based on auditory cues measured on the fly. Velocity inputs of the robot are obtained by modelling the relationship between the variation of given auditory cues and sensors motion. In this way, our approach principally relies on the variation of these cues instead of their intrinsic values. Consequently the localization step is

skipped and the modelling of the acoustic conditions or the effect of the body/head are not required to be modelled. Hence, it is even possible to complete tasks with erroneous auditory cues as long as the variation of these cues remains consistent. Indeed it appears that the variation of auditory cues is more consistent and less sensitive to acoustic conditions than their intrinsic values. This approach applied to robot audition field, is the key contribution of this thesis. The manuscript is thus centred on the description of the AS paradigm that introduces a sensor-based framework, models auditory cues for the latter framework. Experimental studies validate this approach in real world and dynamic configurations either on free field microphones or microphones embedded in humanoid robots. These experiments start from elementary auditory-based task such as head-turn motions, to autonomous navigation tasks.

This manuscript is organized in 6 chapters. Chapter 1 introduces the concept of robot audition. This chapter provides an historical review of the main contributions related to the topic of robot audition. Then, the main applications and contributions, but also the limits of robot audition are summarized.

In Chapter 2, we focus on sound source localization that is one of the most prominent topic in robot audition. As a first step, the localization process is described from human sensing perspective, in order to give an insight of all physiological mechanisms involved in sound localization process. In the following, this chapter gives an overview of localization techniques, and major contributions and applications in robot audition field.

From the limitations underlined in the two previous chapters, in Chapter 3 we introduce the framework that builds up AS paradigm. Before presenting the theoretical concepts involved in AS, we stress the potential benefits of using a sensor-based control through an example comparing an approach based on localization and an approach based on AS. Thereafter, the essential theoretical aspects used in our approach are exposed.

Chapters 4 and 5 that are at the core of this thesis, propose a full modelling and analysis of auditory cues (Interchannel Level Difference, Interchannel Time Difference, Energy level) applied to AS paradigm. These approaches are evaluated through simulations that exhibit the benefits and the limitations of such paradigm. More importantly, all these approaches are validated experimentally on a mobile robot, in real world environments for several tasks and under various contexts. These experiments can be considered as contributions that support the relevance of our approach.

Eventually Chapter 6 emphasizes an essential key point of our approach: the versatility of AS paradigm, that can then be considered as a general framework. More specifically, in this chapter we successfully apply on humanoid robots the frameworks designed in the previous chapters for free-field configurations. Once again experimental results corroborate the pertinence of AS for real world applications.

Chapter 1

Robot audition

This chapter introduces the context of this thesis. Notably we define the concept of robot audition and the related state-of-the-art. The chapter is divided into three main parts.

The first one gives a brief historical review of the progress in robotics and more particularly the developments that conducted to endow robots with auditory capabilities. Following this description, we provide an example of an auditory scene that highlights the needs and expectations from robots endowed with the sense of hearing. And eventually, Section 1.1.3 specifies the functionalities and applications ensued from this sense, shaping the robot audition topic.

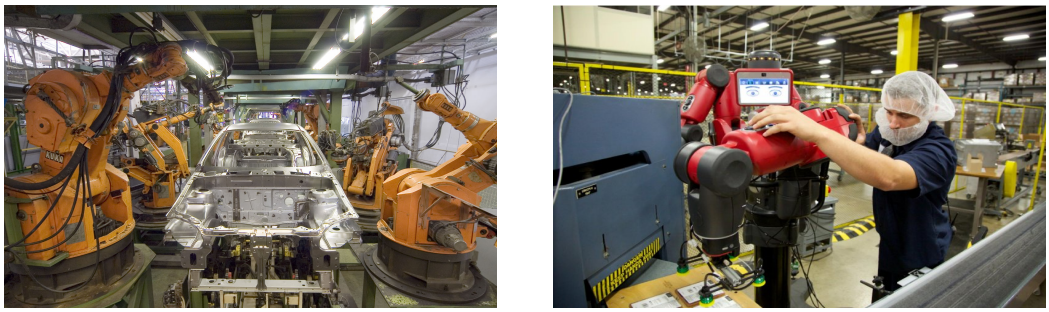
The second part of this chapter concerns robot audition challenges in order to be deployed in realistic environments. First, we focus on the sound propagation properties, which influence the cues extracted from given signals. In a second phase, robotic context is analyzed because of the additional constraints that it introduces.

Finally, the third and last part of this chapter in Section 1.3 presents the current level of development of robot audition through software platforms, research projects and conferences related to this topic.

1.1 The hearing sense in robotics

1.1.1 From robotic arms to robot audition

Since the seventies, the definition of and the expectations from robots have drastically changed. In the beginning of robotics, which coincided with the boom of the industry, robots were considered as efficient tools to expand and develop industrial processes. This period marks the beginning of fully automated tasks designed in order to replace human force for exhausting and tedious works. The first robots were mainly used in controlled environment. Pre-configured robotic arms were programmed for repetitive and tough tasks, such as pick-and-place tasks, palletizing or welding in assembly lines. The environment was static and/or strictly controlled in order to avoid any interruption in working chains and more importantly to avoid accidents in robots working space.



(a) (b)

Figure 1.1: Evolution of industrial robot from the last three decades: in (a) Kuka robots on assembly lines in 1983[1], in (b) Nowadays an assembly line with the robot Baxter (courtesy to Rethink Robotics).

Since that period, robots capabilities have been greatly boosted, although robotic achievements are currently far from the fantasized ideal human-skilled robots. The trigger of this revolution is probably the use of sensors, that made robot more "intelligent" than automated arms. "Intelligent" underlies the concept of advanced skills of sensing, learning and adaptation to unknown and uncontrolled environments. These sensors are from two types and either make robots aware of their internal state or aware of their environment. The internal state of robots is evaluated by proprioceptive sensors that may give measurements about battery level, wheels position, joint position, etc. As proprioceptive sensors one can cite encoders, potentiometers, gyroscopes or compasses. These sensors are particularly useful in industrial contexts in order to detect any defect in robots or breakdown in working chains. On the other part, the environment is observed from exteroceptive sensors. Depending on the context, detecting/analyzing objects or humans and detecting changes in the environment are the main features offered by this type of sensors. Typical exteroceptive sensors include cameras, sonars, lasers, force sensors, microphones, etc. Thus, it is without saying that exteroceptive sensors are greatly used for tasks involving interaction, humans or uncontrolled environments.

Globally, sensors widened the field of application of robotics and democratized the presence of robots in our daily life. In industry, robots tend to be more compliant and to foster human intervention through synergistic tasks (see Figure 1.1b). Nonetheless the use of robots is no more limited to industrial contexts. Indeed robots are more and more involved in tasks such as exploration, monitoring, surveillance, cleaning, socialization or used as leisure companions. These two latter use cases underline the rise of social robots that are autonomously able to interact and to communicate with humans while being endowed with a social behavior. These robots are particularly used in assisting context such as with diseased children [AVS15] or with elderly people [BHR09]. However, in respect of this thesis context, an interesting aspect of robots lies in their nature of compelling platforms for validating and exploring new

[1] By Mixabest CC BY-SA 3.0, commons.wikimedia.org/w/index.php?curid=9820288

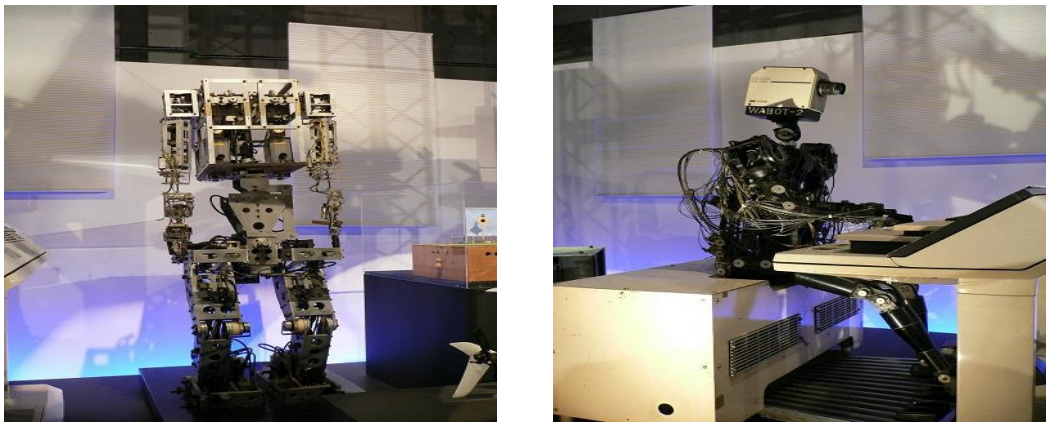
theoretical approaches.

This wide variety of potential applications allowed to diversify robot structures, that are not limited anymore to arms. Indeed, in connection with the sensors that permitted free and autonomous motions, various type of robots emerged: nano/micro robots, wheeled-robots, legged robots, flying robots or humanoids among others. In parallel, with the better understanding of the physiological properties of mammals bodies, senses or behaviors, robots tend to be bio-mimetic platforms. For this purpose, roboticists particularly focus on human-liked robots, that is to say robots endowed with human locomotion capabilities, behavior or sensing skills. The three main senses investigated in robotics are the senses of vision, touch/range sensing and hearing. Among these senses, vision is probably the most prominent and one of oldest subject of study in robotics. Thanks to the astounding progress in electronics, cameras rapidly became embeddable and affordable and subsequently available for robots. On top of that, the support of the long expertise and progress in computer vision field greatly eased knowledge transfer. Robots are nowadays endowed with techniques of detection, recognition, tracking or localization. The sense of touch (*e.g.*, proximity, tactile or force sensing) received more attention as cohabitation between robots and humans increased. The main subjects of study concern collision prevention and detection for safety purpose, grasping tasks that lead to interact with objects of various shape and stiffness, or even locomotion tasks.

Eventually the sense of hearing, that is the main thread of this thesis, consists in endowing the robot with capabilities of analyzing acoustic scenes. Such capabilities require functions for localizing, separating, identifying or recognizing sound sources. Robot audition is then an inter-disciplinary field at the confluence of signal processing, acoustics, AI and control. Robot audition is a quite "recent" topic that has been extensively studied during the last two decades. However, the first trace of auditory capabilities embedded on a robot dates back from 1973 and the first bipedal robot *WABOT-1* [KOK+74] (see Figure 1.2) from Waseda University. This robot was equipped with artificial ears for speech recognition and processing and was able to understand simple Japanese commands such as "STOP", "BEGIN", "TURN LEFT". The following series of robot issued from *WABOT-1* were also all equipped with auditory capabilities. Ten years later, *WABOT-2* [KOS+87] and in the continuity *WASUBOT*, piano-playing robots, were able to converse with a person with a more evolved speech understanding than *WABOT-1*, and to play instruments accordingly to a singing voice from a pitch analysis.

In parallel, the idea of considering robots as validation platforms, especially in neuro-science, greatly contributed to develop and to democratize the sense of hearing for robots in early nineties. In [Web93] and [Web95], the author designed a mobile robot equipped with two microphones to model the behavior of crickets based on sound localization and recognition. Similarly, in [RWE98] a robot head equipped with two microphones and a camera is used to acknowledge the owl neural hearing model proposed by the authors.

As robotics became a subject of research of its own, more contributions were centered on providing robots with auditory capabilities. Auditory capabilities including sound localization and separation based on three microphones were thus proposed



(a) (b)

Figure 1.2: The first robots with auditory capabilities developed at Waseda University: (a) Wabot-1 (1973) a bipedal robot equipped with artificial ears that sustains conversation in Japanese, (b) Wabot-2 (1984) a piano-playing robot that can accompany a human singer.

in [HOS95]. In the middle of the nineties, the authors of [YA95] proposed one of the first robotic system endowed with auditory capabilities for a purpose of human-robot interaction. The robot *Chaser* was equipped with a heat sensor, ultrasound sensors, infrared sensors, touch sensors but also three microphones. By fusing the measurements of all these sensors, this robot was able to locate and to approach a speaker and to start a dialogue thanks to a speech processing module. *Chaser* was mainly designed in order to validate the concept of *active interface* proposed by the authors. In the continuity of theis work, the authors of [SII96] developed a pet robot that is able to localize a speaker from visual and auditory cues by using a neural network. In the same period, [Iri95] proposed to develop auditory capabilities of humanoid robots with a sound localization framework based on neural networks, through the robot *COG*. This latter work was followed by contributions on social robots such as *Hadaly* of [HNK+97] that can localize the speaker as well as recognize speech with a microphone array. Similarly the robot *Jijo-2* introduced in [AHH+97] could understand a phrase command. The auditory capabilities were also extended to navigation tasks in [HSTO97], where the sound localization coupled with sonar sensors allows a mobile robot to approach a sound source while avoiding obstacles. The so-called *cocktail party effect* (see Section 1.1.3), which characterizes the phenomenon of being able to focus the auditory attention on a particular aural stimulus was primarily addressed in [HOS97]. These fundamental researches inspired the birth of the "robot audition" concept.

Lately, with the recent development of autonomous humanoid robots and companion robots, new challenges (see Section 1.2) related to robot hearing emerged. These challenges induced by the autonomy of robots (*i.e.*, flexibility to unknown environments) renewed the interest of robot hearing through the concept of robot

audition, as acknowledged nowadays. In [NOK00b] the authors defined this concept as robots ability to localize, separate, recognize sounds in realistic environment. This definition gave birth to different research topics oriented towards the design of artificial auditory systems. These topics focused on the scattering effect of heads [NMOK03, NOK01] or on the design of artificial auditory systems, such as pinna. For instance, the artificial pinna proposed in [KN11] can move and deform its shape actively in order to improve the sound perception. In the same vein, the authors of [MAB11] or [TR13] studied the best microphones topology to optimize the auditory capabilities. More specifically in [MAB11], the authors proposed a reconfigurable microphone array. Different paradigms of localization were also investigated such as binaural approaches based on two microphones, array-based approaches or more recently *active audition* [NOK00b] that uses the motion. These methods will be detailed in Chapter 2.

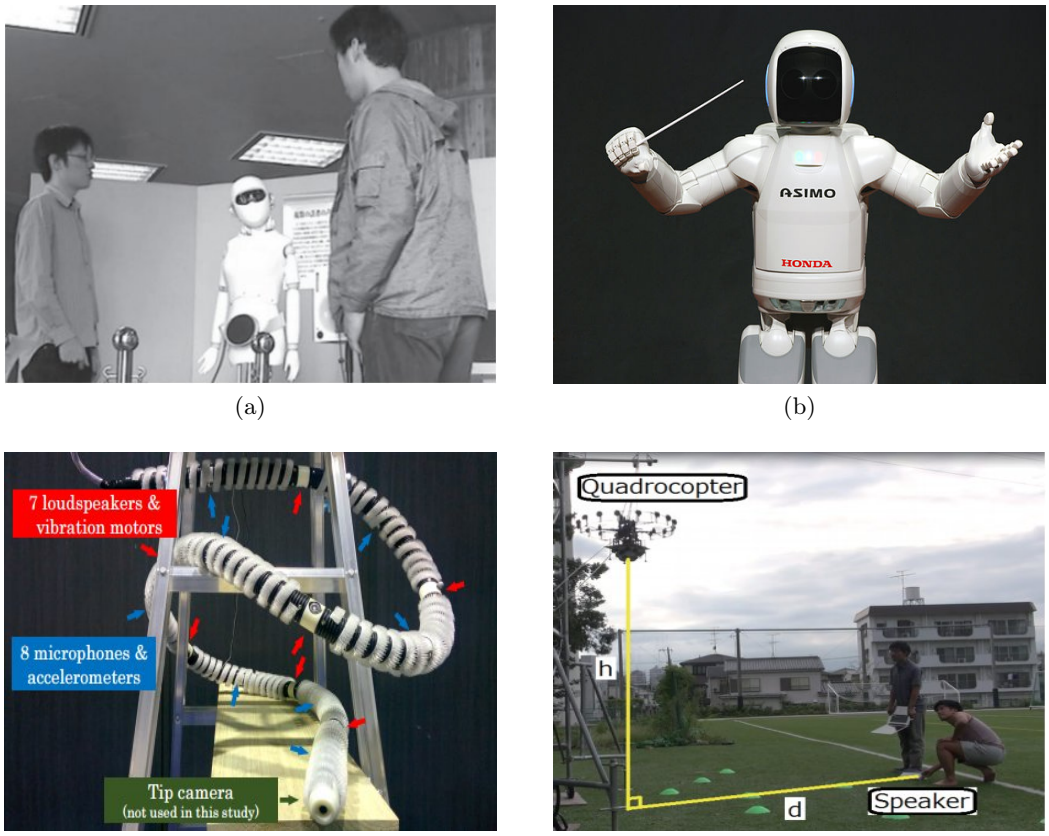
Furthermore, the diversity of robots widened the field of application of robot audition. Robot audition is also a topic related to entertaining robots like dancing robots or music playing robots, as a legacy of the first *WABOT* robots. In this context, the authors of [ANK+99] proposed a theremin-playing robot that uses the audio-feedback related to the sound pitch to control its arm. In [OBI+15], the authors presented a robot that can dance synchronously with music and human dancers from music beats detection. Applications in scenes of disaster is also a recent field of research of robot audition technologies. In [BSLF12] the authors introduced an Unmanned Aerial Vehicle (UAV) system that is able to localize emergency whistles.

This wide variety of applications as illustrated in Figure 1.3 let us foresee that robot audition becomes increasingly important in the field of robotics. However robot audition is still in a growth and maturation phase. Despite a growing interest reflected by the increasing number of projects (see Section 1.3), robot audition community remains modest in size compared to fields like vision. But on the other hand robot audition offers an exciting context that allows ample space for improvements and that promotes original contributions. Indeed most of the applications discussed above are for now limited by the current level of maturity of robot audition: there is still a long way to go before reaching substantially autonomous robots with auditory capabilities.

1.1.2 A typical scenario involving robot audition capabilities

To illustrate the auditory capabilities required for robot audition, let us start from the context of the scene pictured in Figure 1.4. This scene depicts a common everyday life living room in which interactions between humans and a robot are occuring. In this environment the robot should be able to interact and apprehend the surrounding environment with high flexibility. More thoroughly several sensory stimuli may trigger an action from the robot. From an acoustic point of view, several sound sources may be active independently to each other. Domestic noises, such as television, radio, vacuum cleaner, or simply people are each potential sound sources or disturbances. More specifically, in the illustrated scene, the robot action is triggered

[2]By Vanillase -CC BY-SA 3.0, commons.wikimedia.org/w/index.php?curid=17300496



(a) (b) (c) (d)

Figure 1.3: Nowadays a wide variety of application related to robot audition: (a) A human-robot dialogue speech proposed in [TOO14], (b) Asimo leading an orchestra based on tempo and beats detection[2], (c) a hose-shaped robot developed for search-and-rescue mission in narrow and cluttered environment [BIK+15] and (d) sound localization endowed on an aerial robots [ONM+14]

by the boy asking the robot to bring the ball to him. As human, this trivial task is performed efficiently and effortlessly. For robots, this task underlines the need of high-level sensing capabilities. First, an essential question arises: from all the sensory stimuli how to focus the attention on relevant signals and discard "noisy" signals. In our example, the robot is able to segregate sound sources from a signal mixture, and then classify them with respect to their relevance. Furthermore the robot localizes these sources and identifies each of them uniquely: the robot is monitoring acoustically the scene. From this monitoring stage, the robot can determine the location and identity of the boy. The signal content should also be deciphered through speech recognition process. The output of the speech recognition module is then interpreted as an action: the robot heads towards the ball. During robot motion the hearing modules are not disabled. Any new "auditory command" should be interpreted immediately, and the sources of interest are likely to move. A continuous

and real-time sound processing module is running on the robot in order to update the sounds location/presence or the position of the robot with respect to these sources. Furthermore, the motion of the robot generates noise, hence each relevant sound signal monitored is enhanced by denoising and/or dereverberation modules. The ball is finally reached and taken by the robot. Thereafter, the robot heads towards the boy using localization and identification information. This complex sequence of

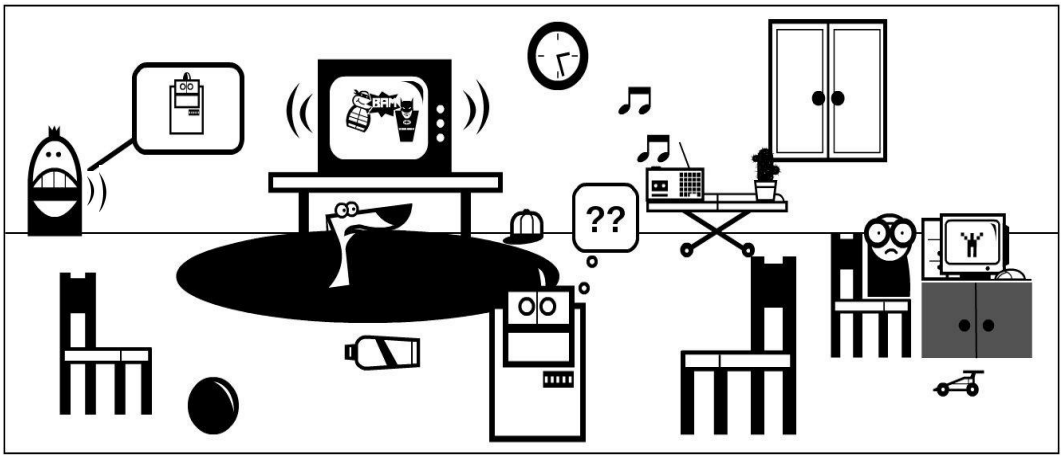


Figure 1.4: A typical acoustic scene to be apprehended autonomously by a robot. Drawing made from (www.stripgenerator.com)

tasks emphasizes complementary capabilities that are based on hearing skills. Endowing each of these capabilities on a robot is a subject of research of its own. The interest in robot audition being quite recent as stated in the previous section, it goes without saying that endowing robots with this kind of hearing sense as imagined in the above scenario, is still an intensive subject of research, that is not achieved yet. In this perspective, more formally, the robot is required to solve questions such as: *Where are the source(s)? What are the sound source(s) of interest? What are they saying? Who are the speakers? When?*. These questions could also be extended to advanced skills that require to solve: *Where am I? What is the shape of the environment? What are the objects?* The capabilities of solving these problems build up robot audition field. These sets of questioning let us foresee two types of auditory-based interaction: the first one concerns the analysis of the sound sources already present in the environment and requires sound localization, sound separation and classification, speech recognition and can potentially involve navigation tasks. The second category consists in analyzing the environment from its acoustic properties and includes methods such as self localization, mapping or scene analysis. These interactions are detailed in the next section.

1.1.3 Hearing skills for robot audition

1.1.3.1 Perception of the sound sources

Where are the source(s)? What are the sound source(s) of interest? What are they saying? Who are the speakers? When?
Solving these issues has been addressed under the theme of computational auditory scene analysis (CASA) in [BC94] or [RO98]. By essence, CASA systems are not specifically designed for robots, but rather consist in a "machine" understanding of an arbitrary sound mixture similarly to human listeners. Typically, CASA is inspired by the psychoacoustic knowledge on human auditory capabilities. Physiologically, human abilities for solving these initial questions have been described in 1953 [Che53] as the *cocktail party effect*. The *cocktail party effect* is an imaged reference to the human ability to selectively focus and recognize a given sound source in a noisy environment, as it could happen in a cocktail party. The hearing interference might be produced by competing sounds or noises that are generally not related to each other. The authors of [HC05] modelled this process with three different neural steps that are

- **Analysis** The first step principally consists in segregating the sound sources with respect to their location. The sound mixture is split so that sources emitted from the same location are grouped together.

- **Recognition** The second step consists in analyzing the statistical structure and patterns in each segregated group obtained from the analysis. These patterns allows to identify a signal of interest among the segregated group.

- **Synthesis** The last step consists in extracting and reconstructing the clean (deconvolved from reverberation) signal of interest from each segregated group. This step provides the capabilities of focusing attention on a signal of particular interest for the listener.

The *cocktail party effect* discussed above gives a quite generic overview on the requirements of CASA for a defined system. Globally the neural process involved in solving the *cocktail party effect* does not necessarily have a precise order and the boundaries between each process is not as neat as it is exposed above. In psychoacoustics there is not yet a global consensus about the neural processing of sound: several studies brought to light experimental evidence supporting different order of processing [Bre94] or [DH97].

In robotics, CASA frameworks are especially centered on human-robot interactions. As exposed in [OOKN04] or [ON15] the interaction processes are generally expressed in a bottom-up framework, depicted in Figure 1.5, that is conceptually easier to develop. In this framework the sound localization is performed first. Then from the location(s) obtained sound separation and selection can be performed if required. The last steps consists in sound identification and processing. The CASA framework emphasizes the prominence of sound source localization that is a prerequisite to guarantee the successful completion of interactions. This framework is slightly

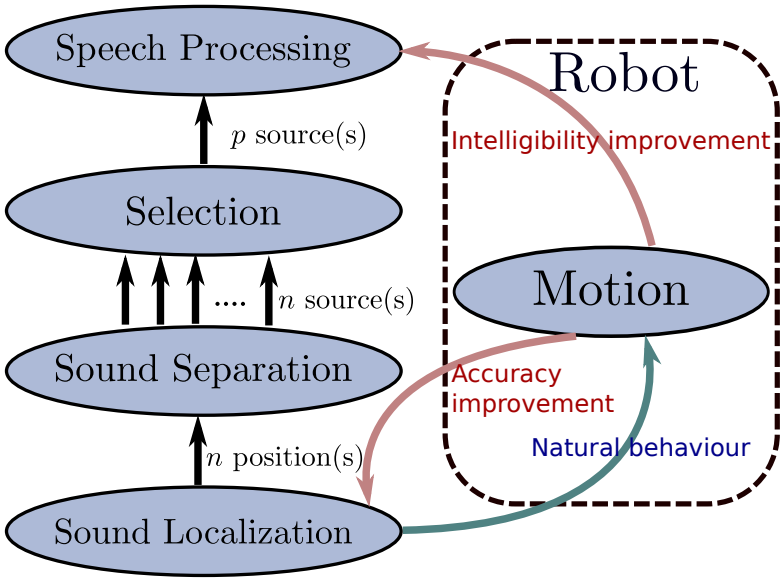


Figure 1.5: Computation auditory scene analysis: the robotic context adds a supplementary element that is the motion of the listener

modified by robotic context that adds an component to CASA: motion. Indeed, the localization process may control robot motions to favor the naturalness of interactions. For instance, it is expected that the robot should face, at a given distance, the speaker with whom the interaction is established. In return, motion may improve the localization with the fusion of measurements from different poses, which corresponds to *active audition* principle [NOK00b]. Speech processing may also be improved by guiding the robot in a pose that optimizes the speech intelligibility. This is applied, for instance, by the authors of [KFKI10] who propose a motion planning strategy, based on occupancy grid, in order to maximize the effectiveness of speech recognition. Nonetheless the motion that can be considered, at first glance, as a valuable addition to CASA framework raises new challenges related to dynamic perception of auditory data, that would require real-time and flexible processing as well as robust tracking (see Section 1.2).

During the last decade, several contributions aimed to apply CASA on robots. These contributions especially focused on sound localization and separation. These approaches evolved from classic techniques such as [AAM99] that is a sound localization and separation method based on a array of microphones to approaches like [DH12], that learns the relationship between the auditory perception and the robot pose. CASA has also been applied for given contexts, as the work [NIYO15] illustrated by Figure 1.6. In this paper, the robot leading a quiz game, is required to perform sound localization and separation, speech processing and identification in presence of multiple speakers that are potentially competing. Similarly, a cooking robot described in [KSN15] is endowed with advanced auditory scene understanding in order to support cooking recipes. Nonetheless these applications of CASA remains far from realistic conditions. Indeed they are generally performed in controlled and

Figure 1.6: Typical CASA application: a robot quiz-master is able to locate, separate and process the sound source in a controlled environment [NIYO15].

static environments. For instance, Figure 1.6 is a relevant illustration of the current limitations of CASA, as floor markers are used to accurately position the robot with respect to the human players.

Robot audition is not limited to human-robot interactions, and several applications may be derived from CASA frameworks. The core statement of robot audition is rather "*robots should understand all sounds picked up by its ear, not just voices*". First, interactions with non-verbal, non-speech sounds might be useful for some particular applications. In amusement culture, music or dancing robots require similar robot audition functions. Dancing robots need to track the beat of musics in order to perform a coherent choreography [YNT+07]. In the case of [ONT+11], a framework allows a robot to play music in collaboration with human performer by analyzing music temporal fluctuation and pitch. These approaches exhibit the strong link between auditory features (*e.g.*, beat, pitch...) and the control of robots. Moreover the sense of hearing is also useful for mobile robotics and brings new navigation modalities based on sound localization. One of the first navigation system based on audition is exposed in [HST+99]. This system consists in a state machine combining the sound localization results to a sonar system for obstacle avoidance, in order to autonomously approach a sound source from a mobile robot. This kind of approach have also recently emerged in the context of distributed multi-robot systems. For instance in [BSLF16] a leader-follower scheme applied is applied to drones in a navigation task based on the localization of the leader.

1.1.3.2 Perception of the environment

Where am I? What is the shape of the environment? What are the objects?
This second set of questioning refers to robots capabilities to sense and analyze their surrounding environments from auditory stimuli. Conversely to source-oriented perception that aims to directly interact with the sources of interest, this section deals

with a higher-level of auditory scene understanding. Indeed, sound from its properties (*e.g.*, acoustic, location, type) can provide information describing acoustically environment structures.

This is the case for mapping and scene reconstruction that can be performed from an auditory perspective. Sound mapping consists in building an auditory representation of the environment, so that zones of interest can be monitored. In the case of mobile robots, sound mapping is developed as an automated tool for condition monitoring, for instance in factories that often contain noisy machines. In this type of environment, any malfunctioning in a machine will be reflected in an acoustic change of the acoustic map. For instance, in [KMIH13] an acoustic map is created while the robot is autonomously navigating through the environment by continuously generating audio scans with a steered response power algorithm (detailed in Chapter 2.3). This kind of approach has also been extended to 3D cases in [EMK+14a]. In this paper, the map is created by a mobile platform equipped with a microphone array and laser range sensors. Similarly, the authors of [EFW+15] proposed a 3D sound mapping using RGB-D sensor to build clouds of 3D auditory points. Globally, these approaches can be considered as extensions of sound source localization and separation techniques: the map is build from the location of the different sound sources. This kind of sound localization and separation remains complex to perform because these approaches should cope with an unknown time-varying number of sound sources that may have different level. These constraints explain that most of these approaches concerning mapping are performed in context of empty scenes. In such environments like factories, sounds produced by the machines are loud, constant, still, and do not vary in number. Sound mapping in populated spaces remains extremely complex, notably because of the non-stationarity of speech, the dynamic nature of the environment and the wide variety of potential sound sources as illustrated in our previous example in Section 1.1.2. Mevertheless auditory scene mapping is of a high interest for interactions, since it would provide auditory awareness ability to robots, that could then focus on or change more easily of interest points. Until now, only few works are interested in this kind of mapping because of the complexity of the task. This problem is addressed in [KGD13] using the cepstral coefficients of sounds to classify acoustic regions. This approach is nonetheless not suitable for real-time implementation and is performed offline. One can also cite the project EARS (see Section 1.3) that targets a demonstration application of a welcoming robot at the reception desk of hotel lobby in 2017.

Other approaches concern self-localization of robots using auditory perception. In this context, sound sources can be considered as acoustic landmarks that can ease robot navigation. One of the first approach of localization was exposed in [WIA04], in which an array of 24 microphones embedded in walls was used to localize a robot. With this system, a robot was able to perform guiding tasks. In connection with search-and-rescue missions, the authors of [BIK+15] proposed a framework to estimate the pose of a hose-shaped robot carrying at the same time several microphones and speakers. In the continuity of approaches aiming to localize the robot, acoustic SLAM (Simultaneous Localization And Mapping) is a recent topic of study. The work exposed in [HCW+11] proposes an acoustic SLAM framework that computes

the robot path using a particle filter and the bearing information of several sources. More recently the authors of [EMN16] developed a framework using the acoustic bearing information on the basis of a probability hypothesis density filter.

In a quite different context, acoustic sensing based on echoes is also a topic of interest. This approach relies on environments acoustic responses with respect to an emitted sound. By analyzing these acoustic responses, information about the surrounding environment can be obtained. This approach can be opposed to acoustic simulator or real-time auralization techniques that estimate auditory sensations from rooms geometrical properties. Auditory responses carry information such as room geometry, material properties, cluttering and potential obstacles. Such approach can be related to mammals like bats or dolphins that use sounds, more particularly echoes, to perceive their surrounding environment. In this topic, [KKK+15] proposed a hammering sound analysis for infrastructure inspection of industrial field based on legged robots. Within this project, a robot can attest of the quality of steel bars from the sound produced by a hammering step. [EMK+14b] used reflections to detect and avoid objects for instance at intersection of corridors, in order to prevent collision with unseen objects. In [SUW15] a robot is able to localize itself in a pre-explored structured environment with odometry measurements and environment acoustic responses. One can eventually cite [KDV16] proposing an echoSLAM framework that can reconstruct the shape of a given environment from wall echoes and potentially retrieve the position of the robot.

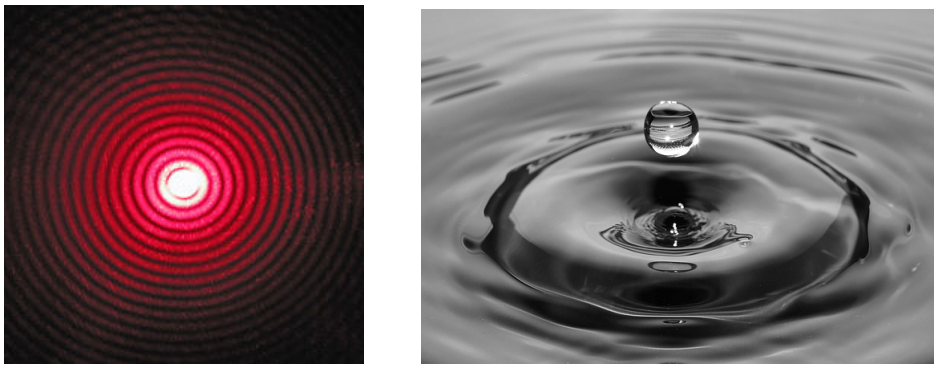
1.2 Challenges in robot audition

The tasks evoked in the previous section, remains challenging to achieve in unknown environments. Until now, there is still a gap between actual auditory capabilities of robots and auditory scenarios such as the one described in Section 1.1.2. For instance, most of human-robot dialogues require the attendant to wear headset microphones or at best to be very close to the robot. In order to characterize the limitations of robot audition in realistic environments, it is important, first, to understand sound properties. Sound as a wave has a particular propagation model and behaviour depending on the environment. From these properties that influence the perception, *i.e.*, the sound heard is different from the sound emitted, the challenges faced in robot audition field naturally arise.

1.2.1 Spherical sound propagation

According to the definition provided by the Acoustical Society of America (ASA) sound can be defined as "*(a) Oscillation in pressure, stress, particle displacement, particle velocity, etc., propagated in a medium with internal forces (e.g., elastic or viscous), or the superposition of such propagated oscillation. (b) Auditory sensation evoked by the oscillation described in (a)*". The definition proposed by ASA covers two distinct but complementary ways to analyze an acoustic scene. In this section, we are more interested in the first definition that considers sound as a physical phenomenon described by wave propagations. This physical phenomenon does not

only describe sound but also the light or fluids like water as illustrated in Figure 1.7. In our context, acoustic theories describe the propagation of a known sound



(a) (b)

Figure 1.7: Example of planar radial wave propagation: (a) in the case of a diffracted laser beam, (b) in the case of water ripple

source through air or any other medium as formulated in [Str77] and [Ray96]. The fundamental results on sound propagation are derived from what is known as the wave equations, that are partial differential equations that govern all patterns of wave propagation. To find a solution to these equations, sound sources are often modelled as vibrating surfaces which have known surface velocities or pressures. One can notice that in a loose sense robot audition corresponds to an acoustic inverse problem: the sound field that is measured by sensors is known while sound sources must be calculated with respect to their location, type or contents. In a homogeneous and linear medium as is the case with air, sound propagation is governed by the following equations

$$\begin{cases} \nabla^2 p(\mathbf{x},t) - \dfrac{1}{c^2}\dfrac{\partial^2}{\partial t^2} p(\mathbf{x},t) = 0. \\ p(\mathbf{x},t_0) = g_1(\mathbf{x}) \\ \lim\limits_{\mathbf{x}\to\mathbf{b_i}} p(\mathbf{x},t) = g_2(\mathbf{x}) \end{cases} \tag{1.1}$$

where $p(\mathbf{x},t)$ is sound pressure at the time t, for the point $\mathbf{x}(x,y,z)$. ∇^2 is the Laplacian expressed in 3D coordinates, and c the constant wave speed inherent to a given homogeneous medium. The initial conditions $p(\mathbf{x},0) = g_1(\mathbf{x})$ and the boundaries conditions $\lim_{\mathbf{x}\to\mathbf{b_i}} p(\mathbf{x},t) = g_2(\mathbf{x})$ with $\mathbf{b_i}$ as boundaries, are required to obtain a unique solution. Note that (1.1) can also be expressed in function of the velocity potential $\phi(\mathbf{x},t)$ or in function of particle velocity $\mathbf{u}(\mathbf{x},t)$, instead of the sound pressure $p(\mathbf{x},t)$. In our context we are particularly interested by the solutions implied by a spherical sound radiation caused by a point source in a free-field (*i.e.*, $\mathbf{b_i} = \infty$). Point sources correspond mathematically to points in space that radiate sound equally in all directions, producing a spherical wave front as illustrated in Figure 1.8. This modelling is interesting since it eases the modelling of reverberations and reflections (see Section 1.2.2.2). Given a source signal $s(t)$ emitted from a position $\mathbf{x_0}$, the wave

equation (1.1) expressed in spherical coordinates [MI68] for the pressure $p(\mathbf{x}, t)$ at the time t is

$$\nabla^2 p(\mathbf{x}, t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} p(\mathbf{x}, t) = -f_{\mathbf{x_0}}(\ell, t). \tag{1.2}$$

$f_{\mathbf{x_0}}(\ell, t)$ is the source function and ℓ the radial distance between the point source $\mathbf{x_0}$ and the observation point $\mathbf{x}$. Considering point sources, $f_{\mathbf{x_0}}(\ell, t)$ characterizes that the sound source occupies a limited and compact region in the space. At each instant t, $f_{\mathbf{x_0}}(\ell, t)$ corresponds to a Dirac impulse so that

$$f(\ell, t) = \delta(\mathbf{x} - \mathbf{x_0})s(t). \tag{1.3}$$

Considering the sound pressure $p(\ell, t)$ at a distance ℓ of $\mathbf{x_0}$ the wave equation becomes

$$\nabla^2 p(\ell, t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} p(\ell, t) = -f_{\mathbf{x_0}}(\ell, t). \tag{1.4}$$

Furthermore, the Laplacian operator is therefore given by

$$\nabla^2 = \frac{1}{\ell^2}\frac{\partial}{\partial \ell}(\ell^2 \frac{\partial}{\partial \ell}), \tag{1.5}$$

assuming spherical coordinates and isotropic propagation. In the frequency domain, the Fourier transform of (1.4) corresponds to the Helmoltz equation[MI68] that is

$$\nabla^2 p(\ell, \omega) + \frac{\omega^2}{c^2} p(\ell, \omega) = -f(\ell, \omega), \tag{1.6}$$

where ω is the angular frequency. The solution of such an equation can be obtained from Green method [Gre28] that transforms (1.6) to

$$\nabla^2 p(\ell, \omega) + \frac{\omega^2}{c^2} g(\ell, \omega) = -\delta(\ell), \tag{1.7}$$

where $g(\ell, \omega)$ is the solution of the equation also called Green's function when (1.3) is satisfied. The Green's function is given by

$$g(\ell, \omega) = K_1 \frac{e^{-i\frac{\omega\ell}{c}}}{\ell} + K_2 \frac{e^{i\frac{\omega\ell}{c}}}{\ell}, \tag{1.8}$$

in which K_1 and K_2 are constants. The general solution can then be written as a convolution

$$p(\ell, \omega) = \int_{-\infty}^{\infty} f(\tau, \omega) g(\ell - \tau, \omega) \mathrm{d}\tau = S(\omega) \left(K_1 \frac{e^{-i\frac{\omega\ell}{c}}}{\ell} + K_2 \frac{e^{i\frac{\omega\ell}{c}}}{\ell} \right). \tag{1.9}$$

From (1.9), it can be observed that the Green's function is a filter that transforms the input signal $s(t)$ into a radiated spherical wave. However as (1.9) is expressed, two solutions emerge under free-field assumption: the first one representing a sound wave radiating from the point source to infinity, and the second one a sound wave radiating from infinity to the point source. From Sommerfeld radiation conditions

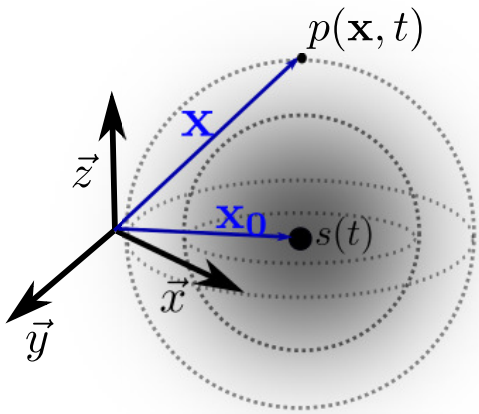


Figure 1.8: The spherical wave propagation in a free-field of a point source $\mathbf{x_0}$ generating a signal $s(t)$ leads to observe a pressure $p(\mathbf{x}, t)$ from a point $\mathbf{x}$ that is function of the delayed and attenuated version of the signal $s(t)$.

[P+81] that state that all sources lie within a bounded region, thus cannot arrives from infinity, it can be deduced the initial conditions $K_1 = \frac{1}{4\pi}$ and $K_2 = 0$. Thus a unique solution can be obtain from 1.9 that becomes

$$p(\ell, \omega) = S(\omega)\frac{e^{-i\frac{\omega\ell}{c}}}{4\pi\ell}. \tag{1.10}$$

Returning in the temporal domain, the latter equation gives the sound pressure as

$$p(\ell, t) = \frac{s(t - \frac{\ell}{c})}{4\pi\ell}. \tag{1.11}$$

Equation (1.11) states that from a point $\mathbf{x}$ located at a distance ℓ of the source, the sound is delayed by $\frac{\ell}{c}$ and attenuated by a factor proportional to ℓ. In Cartesian coordinates the solution of the spherical wave propagation in (1.2) is then

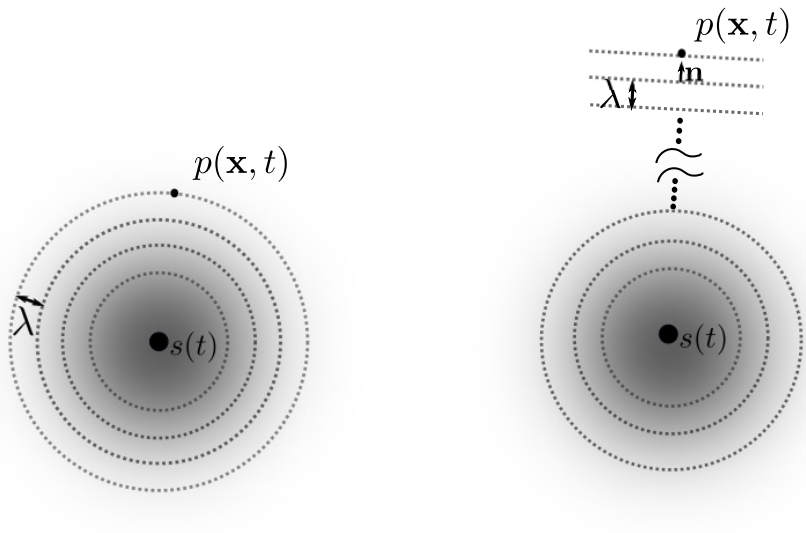
$$p(\mathbf{x}, t) = \frac{s(t - \frac{\|\mathbf{x}-\mathbf{x_0}\|}{c})}{4\pi\|\mathbf{x} - \mathbf{x_0}\|}. \tag{1.12}$$

1.2.2 Environmental constraints

1.2.2.1 Observation distance and plane wave approximation

Now on that the spherical sound wave propagation has been described, one can focus on the properties that such model implies. First it is interesting to analyze how sound behaves at different distance of observation. This is typified by the near-field and far-field sound radiation. If sound pressure is observed at a great distance, the curvature

of the wave fronts looks flat while when the sound pressure is observed at a close distance, the curvature of wave fronts is more pronounced as illustrated in Figure 1.9. The zone of far-field or near-field depends on the observation distance from the point source. Sound is observed from the far-field when in (1.12) $\|\mathbf{x_0}\| \gg \|\mathbf{x}\|$. Conversely the near-field corresponds to observation points where $\|\mathbf{x}-\mathbf{x_0}\| < \lambda$, with



(a) (b)

Figure 1.9: 2D view of the propagation wave with a wavelength λ in (a) the near-field where the wave are curved and in (b) in the far-field where the wave is flat

$\lambda = \frac{c}{f}$ is the wavelength and f the frequency. Note that audible sound has a wide range of wavelength from about 20 mm to 17 m, that corresponds to frequencies of 20 KHz and 20 Hz, respectively. In the field of antennas, the boundary between the far-field (Fraunhofer region) and the near-field (Fresnel region) d_f is generally used to characterize the far field as

$$d_f = \frac{2d^2}{\lambda} \tag{1.13}$$

where d is the inter-microphone distance. This formula is derived from the Rayleigh distance $d_r = \frac{d^2}{2\lambda}$ that characterizes the transition zone from the near-field and the far-field.

In robot audition, observing the source from a far-field region (*i.e.*, with flat front wave) is particularly convenient. Under far-field conditions one can express

$$\|\mathbf{x} - \mathbf{x_0}\| = \|\mathbf{x_0}\| - \epsilon(\mathbf{x}) \tag{1.14}$$

where $\epsilon(\mathbf{x})$ s a residual term. When injecting (1.14) in (1.12), the solution of the wave equation becomes

$$p(\mathbf{x}, t) = \frac{s(t - \frac{\|\mathbf{x_0}\|}{c} + \phi(\mathbf{x}))}{4\pi\|\mathbf{x_0}\| - 4\pi\epsilon(\mathbf{x})} \tag{1.15}$$

where $\phi(\mathbf{x}) = \frac{\epsilon(\mathbf{x})}{c}$ is the residual phase depending on the observation point. Knowing that $\epsilon(\mathbf{x})$ is negligible with respect to $\|\mathbf{x_0}\|$ the latter equation can be simplified as

$$p(\mathbf{x}, t) \propto s(T + \phi(\mathbf{x})). \tag{1.16}$$

where $T = t - \frac{\|\mathbf{x_0}\|}{c}$. Without loss of generality, the proportional gain can be set to 1. From far-field observation properties planar wave fronts have unique directions of propagation driven by a unit vector $\mathbf{n}$. As a consequence $\mathbf{n}.\mathbf{x} = 0$ defines a constant plane where the phase (or time) shift of the signal is the same. Hence (1.16) can be rewritten as

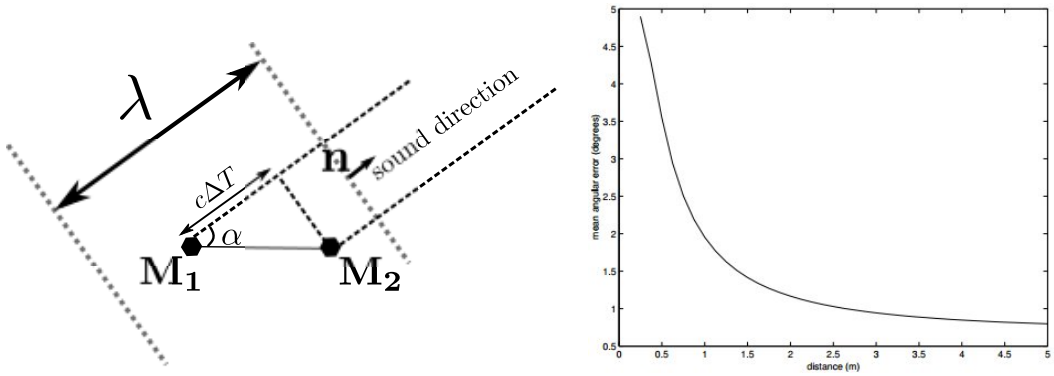
$$p(\mathbf{x}, t) = s(T + \frac{\mathbf{n}.\mathbf{x}}{c}). \tag{1.17}$$

This result states that sound pressure depends only on the direction of propagation $\mathbf{n}$ of wave fronts. This property in robot audition is generally assumed and widely used for sound localization process [NOK01]. In this case the sound source direction can directly be obtained from the time shift from two different observations of the signal as illustrated in Figure 1.10 (see also epipolar geometry in Chapter 2). However as robots or sources location are not supposed to be known, far-field conditions cannot be always ensured. As shown in [VMRL03] and in Figure 1.10, the accuracy of sound localization may drastically decrease as the source get closer to microphones, since the assumption of planar wave fronts does not hold anymore. Hence, some signal processing methods are specially designed to cope with near-field conditions such as localization techniques based on spherical harmonic for MUSIC or beamformer localization for near-field conditions respectively proposed in [KH16] and [ADS06], however with a higher computational cost. Thus having a flexible approach that does not rely on any assumption on the observation field and a limited computational cost is a crucial point for robot audition. Unfortunately the problem is still open, and for now, most of the robot audition solutions consider that the sources are located at more than 1.5 m (which approximates (1.13) for a 8-kHz sound and inter-microphones distance of 15 cm) from the microphones so that the far-field conditions is ensured.

Nonetheless, the far-field condition is not always satisfactory for robot audition. Indeed, because of the sound decay characterized by the attenuation in (1.11) sound localization performance as well as speech recognition performance decreases when the source is too far from the microphones [Rod10]. This is particularly the case when considering indoor environments that are subject to reverberation.

1.2.2.2 Sound reflections

Until now, sound propagation has been modelled in a free space. For a more realistic modelling in indoor environment, propagation models should be discussed in bounded areas. For this purpose, it is required to account for reflective spaces. Rigorously speaking, the following results are dedicated to reflection properties of sound between two fluid media. These results to some extent are applicable to room acoustics, and describe quite well reflections of sound wave by walls. Reflections of sound from solids is a rather complex study, because of the variety of surfaces and the existence



(a) (b)

Figure 1.10: Far-field assumption for the localization: In (a) the direction α of the source is directly linked to the time difference of arrival ΔT of the sound between the microphones $\mathbf{M_1}$ and $\mathbf{M_2}$. Note that the distance between the microphones should be lower than half of the wavelength to avoid 2π ambiguities. However the results in (b) from [VMRL03] show that the accuracy of the localization decreases when the source get closer to the microphones and the far-field assumption is not ensured anymore.

of several modes of wave propagation in solids. For simplicity, let us consider a planar scene implying a plane wave propagating in a medium 1 with a given density ρ_1 and where the sound propagates at a speed c_1. The sound wave reaches a boundary that separates the medium 1 from the medium 2 that has a density ρ_2. The sound in the medium 2 has a speed c_2. The angles θ_i, θ_r and θ_t respectively corresponds to the incident angle, the reflected and transmitted angle of a sound wave as depicted in Figure 1.11.

Sound reflections are governed by the following equations

$$\frac{\sin\theta_i}{c_1} = \frac{\sin\theta_r}{c_1} = \frac{\sin\theta_t}{c_2}. \tag{1.18}$$

From this result it can be inferred that $\theta_i = \theta_r$, which corresponds to the reflection law derived from Fermat's principle in optics, stating that the incident angle of a wave is equal to the reflected angle. Similarly, from (1.18), the Snell's law can be obtain with $\frac{\sin\theta_i}{c_1} = \frac{\sin\theta_t}{c_2}$. The Snell's equation explains the well-known refraction property of sound that is characterized by a change of speed and orientation at the boundary of two media. Moreover the "level" of sound reflection depends on the boundary material. Indeed the coefficients of reflection R_p and transmission T_p are given by the following equations

$$T_p = \frac{2\rho_2 c_2 \cos\theta_i}{\rho_1 c_1 \theta_t + \rho_2 c_2 \theta_i} \tag{1.19}$$

and

$$R_p = \frac{\rho_2 c_2 \cos\theta_i - \rho_1 c_1 \theta_t}{\rho_1 c_1 \theta_t + \rho_2 c_2 \theta_i}. \tag{1.20}$$

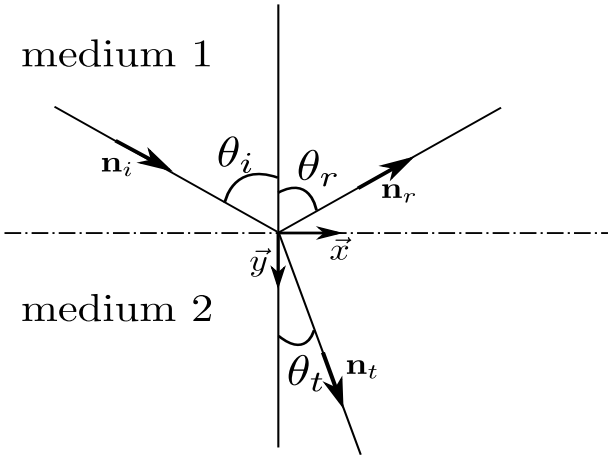
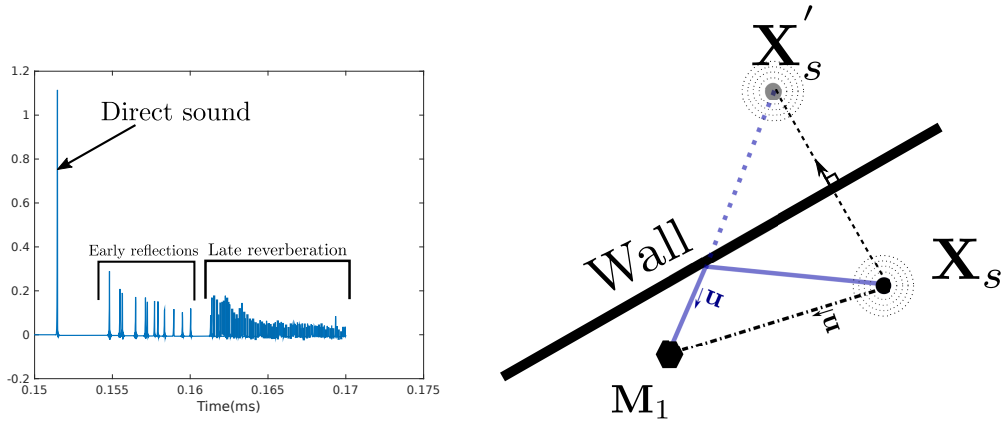


Figure 1.11: A sound reflection at a boundary: the incident sound wave propagating in a medium with $\mathbf{n}_i$ direction and an angle θ_i is reflected at the boundary with an angle θ_r, while a part of the wave is transmitted in the second medium with an angle θ_2

These parameters characterize the level of pressure transmitted or reflected at the boundary of two media. The more similar the two media, the less reflection and the more transmission. This property is particularly used to adjust the acoustic of rooms, that is to say control the level of reverberation depending on the materials of walls. A hard material such as concrete is very dissimilar to the air through which the sound moves, subsequently, most of the sound wave is reflected by walls while fiberglass and acoustic tiles are more similar to air than concrete and thus have a greater ability to absorb sound. A standard measure of reverberation level in a room is the RT_{60} which refers to the time that takes the sound pressure level to decrease by 60 dB. Typical measurements of RT_{60} in indoor environments are given in Table 1.1

In robot audition as in many other tasks in audio processing, sound reflections represent a major issue. The issue lies in the fact that incident waves gets reflected and adds up to the actual sound wave. A geometrical way to represent reflections is to use image-source methods [AB79] by modelling the spherical waves generated by virtual point sources mirrored by an actual source, over each wall (or surface). Hence, as direct and reflected waves start hitting walls, reflections start adding up until they fade out due to sound pressure attenuation (see (1.11)) and walls transmission properties. This phenomenon can be observe from the room impulse response (RIR) that is an acoustic observation of a generated impulse sound. RIR characterizes acoustically a given environment through the direct sound path, early reflections, and the late reverberation as illustrated in Figure 1.12. Thus after several reflections, the sound pressure at the position $\mathbf{x}$ is given by

$$p(\mathbf{x}, t) = \frac{s(t - \frac{\|\mathbf{x}-\mathbf{x_0}\|}{c})}{4\pi\|\mathbf{x} - \mathbf{x_0}\|} + \sum_{r=1}^{\infty} \frac{s(t - \frac{\|\mathbf{x}-\mathbf{x_r}\|}{c})}{4\pi\|\mathbf{x} - \mathbf{x_r}\|} \tag{1.21}$$



(a) (b)

Figure 1.12: Reflection of sound in indoor environment: in (a), the room impulse response characterizing the three modes of sound perception. In (b) the image source model [AB79], expressing that the reflection of a source $\mathbf{X_s}$ recorded by a microphone $\mathbf{M_1}$ can be represented by a symmetric virtual source $\mathbf{X_s'}$ emitting.

where $\mathbf{x_r}$ denotes the virtual point sources. The superposition of all these waves has repercussions on the quality of sound perception. Reverberation corresponds to reminiscences of the actual signal attenuated with a phase shift. The sound waves generated by early echoes are highly correlated and interfere with the direct path signal. Sound waves interference greatly modifies the perceived sound with respect to the actual emitted sound. The observed sound pressure in $\mathbf{x}$ might be greater, lower, or with the same pressure depending on interference phase shifts that condition the way the waves add up (i.e., constructively or destructively). Late reverberation, on the other hand, particularly degrades speech intelligibility. Indeed syllables that are crucial for sound recognition can be completely buried by a long and persistent reverberant field. For instance in Figure 1.13, it becomes computationally challenging to detect the different syllables because of the addition of a persistent sound field slowly decaying, smearing the envelope of the signal.

Furthermore localizing and separating sound sources become more challenging. The strong early reflections illustrated in Figure 1.12 are perceived as potential sources by the localization system. These reflections might even be more perceptible than the direct sound path, especially, when the robot is far from actual sources. Moreover interference might also decrease the pressure of actual sources. Similarly, it becomes more difficult to segregate the sound sources with respect of their location because of interference and the high correlation of early echoes. This motivated the topic of dereverberation that is widely used in speech recognition and by extension could be used in robot audition. The idea is to find a theoretical approach to automatically remove the effect of reverberation from a recorded signal. Reverberation depends on the shape of the room but also on the position of the actual sources, their directivity and eventually microphone(s) position. The solutions considered in

Location	**Volume** (m^3)	**RT**$_{60}$(s)
Recording studio	< 50	0.3
Classroom	< 200	0.4 - 0.6
Private Office	< 1000	0.6 - 0.8
Open Plan Office	< 1000	0.8 – 1.2
Lecture Hall	< 5000	1.0 - 1.5
Concert Hall, Opera	< 20000	1.4 - 2.0
Church	-	2 - 10

Table 1.1: Typical **RT**$_{60}$ in indoor environments from low (green) to high (red) reverberation level. However this classification is based on human sensing. In robot audition 0.3 s is already considered as moderate reverberation while 0.8 s is high.

signal processing literature are mainly based on these information and generally consist in an inverse filtering of room acoustics (see [YSD+12]). For instance [GESS08] and [GNN13] use RIR in order to reduce late reverberation and enhance speech processing stage. A more intuitive solution consists in positioning the robot close to

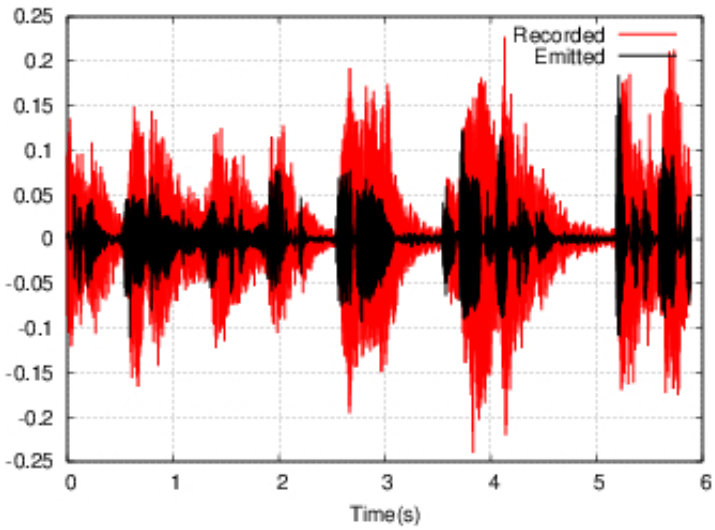


Figure 1.13: Effect of the reverberation on speech: At 3 m from the speaker, in a $4m^2$ room with moderate reverberation, RT60 $\approx$ 400 ms, the speech signal is completely smeared by reverberation which make it computationally challenging to process.

the sound source of interest. Conceptually, it consists in solving the reverberation problem at the sound localization stage. However when facing real reverberant environments the accuracy of localization techniques remains averaged. Localization under conditions of high reverberation still remains a challenging task and most of the techniques are very sensitive to variations of acoustic environments. For example, the authors of [VMR07] proposed to subtract reverberation spectrally similarly to noise

spectral suppression techniques (see Section 1.2.3.2). This approach requires some tuning parameters that depends on room acoustics. Otherwise, the main path of improvements relies on enhancing the GCC angular spectrum with spectral weighting function such as the so-called PHAse Transform (PHAT) [KMOO11, LLY11] that is known (heuristically) to be more robust to reverberation. Other methods try to benefit from the redundancy of measurements by physically increasing the number of microphones or through *active audition* framework [NOK00b].

1.2.3 Challenges raised by robotic context

1.2.3.1 Technical constraints

The first challenges raised by robotic context concern the technical and physical limitations inherent to this type of machines. These limitations are unique compared to other hearing machine systems. First, robots are machines that meet physical and geometrical constraints due to their structure. An auditory system should be embeddable on a robot considering space restriction. Space limitation and embeddability of auditory systems are conflicting with the purpose of robustness and accuracy. Indeed "hearing" performance can be improved either by using a maximum number of microphones in order to increase the redundancy of the measurements or by maximizing the overall array aperture, namely the distance between microphones as stated in [Tre02]. A maximal distance between the microphones improves the accuracy of the localization since the directivity of the source gets smaller as the spacing between the microphones increases. But on the other hand, as the spacing between the microphones increases, spectral aliasing is more likely to happen for high frequencies and the far-field zone moves away (see (1.13)). The far-field assumption usually admitted in sound localization becomes more difficult to ensure. From the latter results, optimal positioning and topology of transducers (such as microphones) build up a thorough subject of research studied in antenna and microphone arrays literature. Some contributions focus on optimizing array configurations in order to increase the number of detectable source such as the solutions proposed in [PPL90]. In [ZB10], the authors introduced an optimal array configuration that improves the localization accuracy by maximizing the mutual information between the actual source location and the estimated one. These techniques are unfortunately not applicable on real robots due to the embeddability constraint among others. For instance, on humanoid robots the space between microphones does not exceed few tens of centimeters in general. A tradeoff between performance, accuracy (*i.e.* far-field assumption) and embeddability of auditory systems is required: commonly, on robots the number of microphones is limited to eight, for microphone array systems, while binaural systems are limited to two microphones. Hence, in this case, robust algorithms should be developed with very few and close microphones.

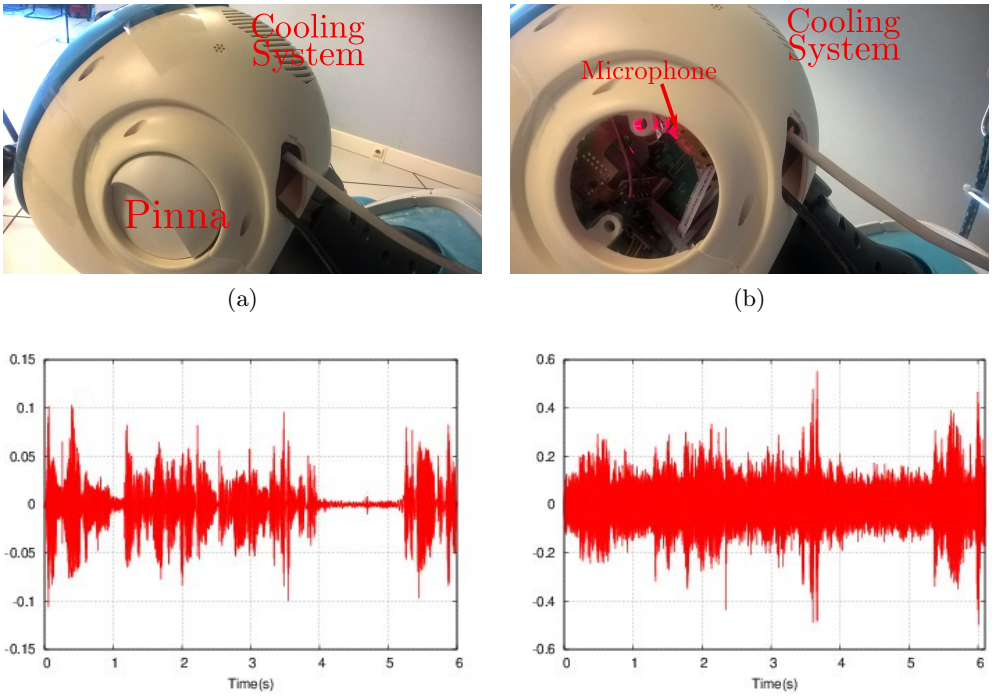
Another aspect of the embeddability constraint concerns real-time requirements that entail restrictions on available computational resources and the maximum payload of sound processing algorithms. As a consequence auditory approaches for robots should be low demanding on resources, which limits their complexity and their computational cost. For instance, sound localization estimations must be avail-

able within a short time interval. Typically, up to 150 ms are acceptable for reflex tasks such as sensor-based control. Inverse methods developed in signal processing such as inverse boundary-element models [AT00], basically localization techniques relying on acoustic wave equations described in the previous section, are not suitable for robot audition because of their computational cost. Furthermore real-time processing implies to limit the signal duration from which the localization is performed. The signal duration is generally less than 100 ms for real-time localization. A short signal duration decreases localization performance as stated in [VVO04] for humans and in [MMWB15] for machines. Similar real-time constraints hold for speech processing algorithms in order to guarantee an appropriate social behavior during interactions. Fortunately, in this case, this constraint is generally fulfilled thanks to the recent progress and wide application of automatic speech recognition for machines (*e.g.*, mobile phones) that also requires real-time computations. The main challenges related to automatic speech recognition in robotics are about the perturbations caused by the environment as discussed previously and ego-noises that are discussed in the subsequent section.

At last, a fundamental difference between robot audition and classic machine hearing systems, is obviously motion: robots are moving and evolving autonomously in dynamic environments. Auditory techniques should then be able to perform dynamic perception and to cope with acoustic changes in the environment. These constraints are challenging to solve and until now only few works have tried to tackle this problem in robot audition. The main obstacles come from auditory cues that are continuously changing during data acquisition. Consequently inconsistent cues are likely to be extracted and may degrade the hearing capabilities of the robot. Generally most approaches perceive sound in a static way as previously mentioned with CASA applications in robotics [NIYO15]. When robots are moving, the path taken by many approaches consists in the "stop-and-perceive" principle that suggests stopping the robot while acquiring data [NOK00b][TTLJ15]. This kind of approach adds latency and may prevent robots from receiving new commands besides altering interactions with saccadic behaviours. Recently in [TR15], the authors proposed a method to improve localization cues by compensating the motion of the robot. However this method assumes anechoic conditions. Some other approaches address this situation in an opposite way: a fixed auditory system with a moving sound source, which can be considered conceptually as the same problem. For localization, [PDA12] introduced a stochastic approach to cope with a moving sound sound source, that has been validated in anechoic conditions. Tracking systems based on particle filters as proposed in [VMR07] or probability hypothesis density trackers [EMN+15] are also alternative solutions to cope with moving sources. Similarly in [NNIH10] or [NNHT09] the authors proposed a sound source separation and recognition technique for moving sources based on an adaptive windows size. Nevertheless, most of the aforementioned works are performed either in simulation or in ideal acoustic conditions (*i.e.*, anechoic or prepared room) and with an array of microphones. The degree of complexity raises to another scale when in addition to the motion of the robot, the source of interest is moving, as it may occurs in realistic scenarios.

1.2.3.2 Ego-noise

Another challenge consists in dealing with ego-noises, that is to say noises caused internally by robots. Noises can be caused by several type of perturbations. One special type of ego-noise, which is observed while robots are performing actions using their motors, is called ego-motion noise. Indeed joint mechanical frictions, quite present in humanoid robots, or motors that control robot motions increase the level of noise in the recorded sound signal. A different type of noise is caused by the embedded electronic systems, that generally use a cooling system with loud fans generating noisy perturbation. Eventually noises can be caused by the utterance of robots during barge-in situations: humans could freely speak during robots utterance requiring to "clean" recorded signal.



(a) (b)

(c) Sound played (d) Sound recorded

Figure 1.14: An example of ego-noise caused by an unsuitable microphone configuration: the sound played in (c) from a loudspeaker is severely altered by the cooling system noise as illustrated in (d).

One can deal with ego-noises at different level. First, they can be limited at a mechanical and topological level, which refers to robots conception stage. Microphones can be positioned in such way that they are far from potential sources of noise such as joints and cooling systems. Nonetheless, this solution is quite restrictive and is limited by the type of robots considered. In this context robots manufacturers play a major role in robot audition development. Indeed as the interest for audition appeared quite lately in the timeline of robotics, the mechanical conception

of robots regarding the auditory system efficiency was not, until recently, a priority. Thus it is not surprising that many commercial robots do not embed auditory systems or possess auditory systems located at unsuitable positions (*e.g*, near fans) that severely degrade sound quality as illustrated in Figure 1.14. Such noisy signals may seriously deteriorate localization and speech processing capabilities of robots. Fortunately, with the higher demand of social robots coupled with the increase of robot audition projects (see Section 1.3), specific robots designed for or supporting robot audition are arising. On can refer to the robotized head *POPeye* [CBL+09] or the dummy head *KEMAR* [BDFP16] that are specially designed for robot audition research. Moreover, recent projects such as EARS, involving robot manufacturers (*i.e.*, Softbank Robotics) participate to raise the awareness about the importance of designing consistent microphones topology.

A second solution to deal with ego-noises is to use spectral subtraction algorithms that are particularly suitable for stationary noise. The denoising filters proposed in the literature can be integrated in hearing process of robots. This is the case of the well-known like Ephraim-malah Filter [EM85] that consists in filtering spectrally the noisy signal with respect to an estimate of the signal-to-noise ratio (SNR). To do so, let us consider the Short Time Fourier Transform (STFT) $X(k, f)$ of the noisy input signal $x(t)$, where f and k respectively stand for the frequency and the short-time frame. The Ephraim-Malah spectral gain $G(k, f)$ is then defined as:

$$\begin{aligned} G(k,f) =& \frac{\sqrt{\pi}}{2}\sqrt{\frac{1}{1+SNR_{post}(k,f)}\frac{SNR_{prio}(k,f)}{1+SNR_{prio}(k,f)}} \\ &\times \mathbf{B}\left[(1+SNR_{post}(k,f))\frac{SNR_{prio}(k,f)}{1+SNR_{prio}(k,f)}\right] \end{aligned} \tag{1.22}$$

where the function $\mathbf{B}$ is the function

$$\mathbf{B}(x) = e(-x/2)\left[(1+x)I_0(x/2) + xI_1(x/2)\right] \tag{1.23}$$

in which I_0 and I_1 stand for the Bessel functions of zero and first order. Equation (1.22) depends on $SNR_{prio}(k, f)$ and $SNR_{post}(k, f)$ that are respectively the *a-posteriori* and *a-priori* SNR. $SNR_{post}(k, f)$ is given by

$$SNR_{post}(k,f) = \frac{|X(k,f)|^2}{|N(f)|^2} - 1 \tag{1.24}$$

where $|N(f)|^2$ denotes the noise spectrum. $|N(f)|$ can be easily obtained from a sample of signal that contains only noise or in a more accurate way from the Minima-Controlled Recursive Average (MCRA) algorithm [CB02]. The MCRA algorithm estimates the noise spectrum by averaging past spectral power values with respect to the speech presence probability. Thence, $SNR_{post}(k, f)$ is an estimate of the SNR of the current short-time frame. On the other hand, $SNR_{prio}(k, f)$ is the prediction of the ratio of the power of the clean signal and of the noise power that is recursively estimated from decision-directed approach proposed in [EM85] as

$$SNR_{prio}(k,f) = (1-\alpha_a)\mathbf{P}[SNR_{post}(k,f)] + \alpha_a\frac{|G(k-1,f)X(k-1,f)|^2}{|N(f)|^2} \tag{1.25}$$

where $\mathbf{P}(x) = x$ if $x > 0$ and 0 otherwise, and $0 < \alpha_a < 1$ is a smoothing parameter. Some other denoising filters exist in literature such as Wiener filters that have been used in robot audition in [VMR07] or [MP10]. In these papers, the spectral weighting $G(k, f)$ function is simply defined as

$$G(k, f) = \frac{SNR_{prio}(k, f)}{1 + SNR_{prio}(k, f)}. \tag{1.26}$$

More recently in [GM15], the authors proposed to improve $G(k, f)$ by considering (1.26) as a hard mask that is set to 0 or 1 depending on the value of $SNR_{post}(k, f)$ with respect to a given threshold. In [GM16], they enhance the noise filter $G(k, f)$ with a transition function that can cope with abrupt changes in broadband noise level by relying on the non-stationarity of the speech compared to the noise. Globally this kind of approach is particularly efficient for stationary and uncorrelated noise since used a sample of noise $|N(f)|$ that is not suppose to vary. Hence, in robot audition, ego-noise caused by fans for instance, can be reduced with such techniques.

However the problem remains open for internal noises caused by motors or joint frictions of the robot. Unlike sounds emanating from the cooling systems, the noise level emanating from motors is unpredictable, highly fluctuating and in some cases directional. For instance in UAV, the sound produced by the rotors and propellers is particularly loud and vary with the motion and the speed of the vehicle. The problem has been particularly addressed in the context of sound source localization which performance is strongly degraded by noise level. In robot audition literature, a first type of approach proposes to perform sound separation by splitting internal sounds from external sounds. In [NOK00a] the ego-motion noise is measured by an internal pair of microphones in order to assess the relevance of the measurements of the external microphones. In [ESS+09] the authors use blind sources separation processes to split the internal sound from the external thanks to additional sensors that only measures the internal noise. Nevertheless this kind of approach requires additional internal sensors, which is not possible on every robot. And more importantly the efficiency of these methods depends on the strong assumption that the internal sensors measure only internal noises.

The second type of approach considers the mapping between the motion of the robot with the noise generated. In [INR+09] and [INA+11], the method of mapping is based on template matching techniques with the idea that motor noises can be predicted, if the motion performed by the robot has a pattern of a prior known duration. Predicted noises are then subtracted from the noisy signal. Similarly, [IKSM05] introduces an approach based on neural networks that predicts the noise spectrum from angular velocities of the joints of the robot, while in [NNN+06] the ego-motion noise is subtracted using templates recorded in advance for each motion and gesture involving activity of several motors at a time. Recently, the authors of [DK15] proposed to create a dictionary by combining a K-SVD [AEB06] and the sources phase in order to extract a learned signal from a new input mixture. This approach has been applied to ego-noise reduction. The noise related to each motion of the robot could be learned and extracted afterwards from the microphones recordings. These approaches are limited by the potential high number of joints of

the robot: for complex and dynamic robots, the prediction of the ego-noise becomes particularly challenging. Generally ego-motion noise predictions are limited to few motions such as in [NNA+09], that focus on the noise caused by the robot head rotation, or for UAVs, that are underactuated systems (*i.e*, simple motions compared to anthroppomorphic robots), as introduced in [FON+13]. The latter paper uses an adaptive estimation method of the noise spectrum from motor control values and the state of the UAV.

1.3 Robot audition research

1.3.1 Research projects

Globally, the previous section attested of the main difficulties of robot audition when facing real-world environments. To face these challenges, various research projects were issued from collaborations of several research teams during the last decade. These fruitful collaborations impulse the path taken by robot audition community. In this section, we review these projects through their achievements.

1.3.1.1 Software and hardware solutions

To begin with, a description of the software platforms available in robot audition is important. Currently there are two major softwares for robot audition that are detailed below.

HARK[3]: Based on their early contributions on robot audition for humanoid platform like *SIG*, the authors of [NMOT04, NOK00b] released a open-sourced robot audition software HARK (Honda Research Institute Japan Audition for Robots with Kyoto University). HARK is the result of the 5-year research project **Robot Audition from Computational Auditory Scene Analysis** started in 2007. As outlined by HARK structure, this project focused mainly on the elementary function to build up a CASA framework, that is to say, binaural [KMOO11] or array-based [ONOO11] localization, sources separation [HYT+12] and speech processing [SKHO08]. HARK consists in a set of module corresponding to the latter functionalities to construct a robot audition framework independently of the microphones configuration or the type of robot used. Currently the project **Multiple Development of Robot Audition** aims to extend HARK features to cope with moving sound sources and/or robots [NNA+09] and the development of robot audition on aerial vehicles [OYNN12]. More recently the project **Impact**, aims to extend the application of robot audition in the context of search-and-rescue missions, and advanced functionalities are expected to be implemented in the software. The details of the techniques and performances for source localization, separation or speech recognition are given in [NTO+10] or more recently in [ON15]. Globally, the aim of HARK is to provide a modular common platform for robot audition community with high usability and real-time performances. The software is regularly updated and sup-

[3] http://www.hark.jp

ports several library and tools such as ROS (Robot operating System), OpenCV, or Kinect.

ManyEars[4]: The second software, the ManyEars project is an open GPL software principally based on the work published as [VMR07]. ManyEars [GLF+13] provides real-time sound source localization and separation, tracking for several sound sources using an array of eight microphones. ManyEars is implemented in C as a modular library, with no dependence on external libraries. Attached to this project, the 8SoundUSB hardware [AC+] provides an interesting sound acquisition system that consists in a dedicated sound card and eight microphones. This solution can be especially useful for mobile robot, since only few dedicated robot audition hardware exist. The commercially available sound cards and microphones are not designed to be embedded on robots and may integrate functionalities that are not useful for robot audition (*e.g.*, mixing, sound effects...). Eventually, similarly to HARK, the software is also compatible with ROS.

1.3.1.2 Collaborative projects

Research projects concerning robot audition has significantly increased in the last few years, showing the growing importance of robot audition in the scientific community. We thus review below the main projects related to robot audition that we are aware of.

POP: POP (Perception on Purpose) was a 3-year scientific project that lasted from 2006 to 2009. The project concentrated onto the understanding and modeling of the interactions between an agent and its physical environment. These interactions were modelled from auditory and visual cues, with a biological and computational perspective. The main topic concerned the models of cognitive mechanism of attention, the extraction of auditory and visual cues as well as their integration and the study of coordination between the sensor observations and the motor activities. Although the studies mentioned in the project were not especially designed for robotic context, a binaural robotized test platform, *POPeye* [CBL+09], has been developed in order to validate the theoretical findings from the aforementioned researches. From robot audition viewpoint, this project developed novel algorithms for real-time robust localization of sound sources [CMWB09] based on pitch and interaural cues. The project also thoroughly studied the active perception of sound from which one can improve the sound localization performance [CLLH07] by turning the head in the direction of the source, or recover the distance to the sound source [LC08]. Furthermore the database CAVA[5](Computational Audio-Visual Analysis) was released in order to develop and evaluate computational methods of audio-visual scene analysis. Eventually, POP involved partners that are Inria Grenoble, more particularly *Perception* team, the University of Sheffield, through the *Speech and hearing* group, the University of Osnabrück with *Institute of Cognitive Science* group, the University Hospital Hamburg-Eppendorf *Institute of Neurophysiology and Pathophysiology* and the *Institute for Systems and Robotics* of Coimbra University.

[4] https://sourceforge.net/projects/manyears
[5] http://perception.inrialpes.fr/CAVA_Dataset/Site/

HUMAVIPS: In the continuity of POP, HUMAVIPS (Humanoids with Auditory and Visual Abilities In Populated Spaces) was also a 3-year project started in 2010. HUMAVIPS was particularly centered on audio-visual perception for Human–robot interaction and humanoid behavior. The objective of HUMAVIPS has been to endow humanoid robots with audiovisual abilities that is to say to explore a populated space, localize people and possibly engage interaction with one or two people with an appropriate behavior. The main idea of this project was to build-up the tools in order to solve the *cocktail party problem* with the help of visual sense. This project led to contributions such as sound event recognition [JAPGH12] or categorization [APSRH13], source localization and separation through learning methods [DH12] or method of sound localization based on geometrical time-delay [APH12]. Similarly to POP, a database of binaural recordings CAMIL[6] (Computational Audio-Motor Integration through Learning) provides real-word recordings of various type of sources (speech, music, noise...) and configurations (still or moving recording system). This project involved *Perception* and *Mitis* team of INRIA Grenoble, Aldebaran Robotics (currently known as Softbank Robotics), the *Research Institute for Cognition and Robotics* at Bielefeld University, IDIAP Research Institute and the *Center for Machine Perception* at Czech Technical University.

BINAAHR: BINAAHR (Binaural Active Audition for Humanoid Robots) project was established from 2009 to 2013 as a French–Japanese collaboration focused on active binaural robot audition. The project aimed to endow humanoid robots with auditory capabilities inspired from human auditory modelling. This project investigated the relationship between motion and auditory perception through *active audition* with contribution such as [PDA12] or sensorimotor theories for auditory perception in [BPDCG12]. The multi-party communication was also a topic of interest of the project with high level functions such as speaker identification [YBA+11], speech recognition with audiovisual activity detection [YAZ12a], attention mechanisms [YTK+11] [HTOO11], sources separation [NTOO12], ego-noises and reverberation cancellation [TNT+12]. BINAARHR also included software development with several of the aforementioned contributions which were incorporated to HARK, and hardware development with a design of an artificial binaural sensor based on EAR sensor, a fully programmable eight-channel data-acquisition board [BADM09]. Finally this collaboration involved LAAS, ISIR and LPP from French side, while Kyoto University, Tokyo Institute of Technology (TITECH) and Kumamoto University were involved in Japan.

TWO!EARS[7]: This project started in 2013 and is ending by 2016. The aim of this research project is to develop a computational framework for modelling active human listening that enables the analysis of auditory scenes. The main difference of this project with respect to the usual auditory framework lies in the two-way process of the modelling. A bottom-up framework as described in Figure 1.5 is used for the acoustic scene understanding, but also a top-down process where hypothesis from the acoustic scene are integrated in order to improve the elementary functions of sound localization or separation [SWK+14]. Although this project is devoted to the wide

[6] http://perception.inrialpes.fr/~Deleforge/CAMIL_Dataset
[7] http://twoears.aipa.tu-berlin.de

field of machine hearing, several works have been validated on robotic platforms. This typical approach is used for instance in [PBDM14] that use prior independent learning of the interaural transfer functions and the environment noise statistics to develop a sound source localization method. *Active audition* is also a major topic of the project with contributions that use motion to improve the sound source localization [MMB15] or use feedback loop [BDFP16] to minimize the localization uncertainty. Learning methods are also explored with deep neural networks for localization [MBM15] or speech recognition [MMBB15]. As a result of this project, an auditory framework was released as a GNU software that includes several stages for modelling active human listening (localization, separation, speech processing...) coupled with a blackboard architecture that allows the inclusion of world knowledge, as well as an acoustical simulation framework for creating binaural signals. TWO!EARS includes several partners such as LAAS, ISIR, the University of Sheffield, through the *Speech and hearing* group, the Institute of Communication Acoustics, Ruhr University Bochum, the Department of Electrical Engineering-Hearing Systems from Technical University of Denmark and others.

EARS[8]: EARS (Embodied Audition for RobotS) that will last until 2017 is centered on human-robot interaction in adverse environment for a technical solution that should be embedded on a commercial robot (NAO from Softbank robotics). The solution will be evaluated through a final demonstration in a scenario of a welcoming robot at a reception desk. For this perspective, EARS focus on designing an efficient and embeddable microphone array [TR14] and sound localization techniques devoted to this configuration like [EMN14] or [MEN+15]. A part of the project is also devoted to scene analysis through audio and visual modalities via [GAPFH16] or [GBEH15]. For robustness and high applicability, EARS project is based on array-based processing with a particular focus on dynamic scenes that cover localization of moving sound sources [EMN+15] and sound separation of moving sources [KBGAP+15] or localization while the robot is moving [TR15]. The project involves an industrial partner (SoftBank Robotics), and scientific partners including *Perception* from INRIA Grenoble, the *Speech and Audio Processing Group* of Imperial College London, the *Multimedia Communications and Signal Processing* group from Friedrich-Alexander-University, the *Acoustics Lab* from Ben-Gurion University and *Adaptive Systems* of Humboldt University of Berlin.

1.3.2 Conferences and organized sessions

Since 2000s, the growing interest of the scientific community in robot audition has also been reflected in the increasing space given to audition in international conferences. These conferences are opportunities to raise awareness of robot audition issues among the community, and by the same mean to increase the interest towards such a challenging topic.

First of all, among robotics conferences, IROS (International Conference on Intelligent Robots and Systems) and ICRA (International Conference on Robotics and Automation) are probably the biggest roboticists meetings and give a relevant

[8] https://robot-ears.eu

overview of the current state-of-the-art of robotics in general. The first sessions about robot audition appeared in IROS 2003. And until 2013, sessions about Robot Audition were organized at the instigation of K. Nakadai and H. Okuno. Nowadays, since 2014, "Robot Audition" is registered in the keyword list of research topics of RAS (Robotic and Automation Society), and regular sessions on the topic are organized at IROS each year. IROS is probably the conference that gathers the most contributions related to robot audition each year. For instance in IROS 2015, 26 papers with the keyword "Robot Audition" were submitted. In comparison with computer vision field which account for 78 papers, this number is relatively low but is nevertheless in the head of list of frequent keywords. Among these papers, 14 papers were accepted and presented at the conference. Moreover regularly special sessions are organized at IROS in order to synthesize the main progress in the field and set roadmaps and direction of future researches. For IROS 2016, the special session "New Horizon for Robot Audition Application" organized by K. Nakadai, H. Okuno and others is centered on applications based on robot audition technologies, and innovative technologies towards such applications. Regarding ICRA, 2015 edition saw the inclusion of the keyword "Robot Audition", while this keyword was surprisingly missing for ICRA 2016 edition. Until now, there is no regular session devoted to robot audition in this conference. As a result, generally few papers concerning robot audition are submitted to ICRA. For instance in ICRA 2016, only 4 papers related to robot audition were presented, dispatched in diverse sessions. Robot audition contributions can also been found in different robotic conferences. One can cite RO-MAN (International Symposium on Robot and Human Interactive Communication), RO-BIO (International Conference on Robotics and Biomimetics) or HRI (International Human-Robot Interaction Conference), where robot audition applications are used in the context of the latter conferences, that is to say for interactions or as cognitive processes.

In signal processing, ICASSP (International Conference on Acoustics Speech and Signal Processing) one of the biggest conference in signal processing sometimes holds special sessions about robot audition. In the last decade, a handful of robot audition sessions were organized, such as "Signal Processing Techniques and Algorithms on Robot Audition" organized by A. Sugiyama and H. Okuno in 2009 or “Audio for Robots – Robots for Audio” in 2015 organized by E. Vincent and J. Le Roux. However most of the contributions in ICASSP are generally oriented towards theoretical aspects of signal processing unlike robotic conferences that encourage applications. Targeting this conference is also a good way to encourage signal processing community to consider robotic context for their developments or at least real-world applications. In parallel other conferences related to signal processing involve the participation of robot audition community. For instance EUSIPCO 2015 (European Signal Processing Conference) held a tutorial about EARS project, organized by H. Löllmann, C. Evers and R. Horaud. One can also cite INTERSPEECH (Conference of the International Speech Communication Association) or ICA (International Congress on Acoustics) that are however more centered on computational machine understanding of auditory perception.

1.4 Conclusion

To sum up this introductory chapter, robot audition topic represents a promising path of improvement of robots capabilities beside their interactions with the environment. Hearing capabilities complement vision and touch senses for robots interacting in populated spaces. Thus, endowing robots with a reliable hearing sense is a crucial path for scene analysis and would allow robots to apprehend more comfortably realistic environments. The main needs in robot audition concern sound localization, sound separation and speech processing. These elementary functionalities build up most of the applications in robot audition. Robot audition can be used in a wide variety of application such as human-robot interaction, navigation, search-and-rescue missions or mapping without restrictions on the type of robot (humanoids, UAV, mobile robots...). However despite this tremendous potential, there are little research in robot audition in comparison to the huge amount of research on robot vision, for instance.

The explanation lies in the fact that robotic audition is extremely complex and limited especially in real-world environments. First the limitations come from the inherent properties of sound waves, that induce interference from reverberation or modelling constraints on the field of observation of the sound. Moreover, environments that may be dynamic and may include several types of sound sources, noise, raise the degree of complexity of auditory perception. And eventually the requirements induced by the robotic context, such as the real-time constraints end up making robot audition a very challenging topic. Modelling accurately all the potential perturbations and environment variability turns out to be complicated, and many realistic situations remain unsolved until now. As a result, most of the models and techniques proposed in the robot audition literature consider static scene, that are either validated in simulation or experimentally in controlled conditions such as anechoic rooms.

To tackle the robot audition limits, sound source localization, among the functionalities related to robot audition, is certainly the subject that receives the most attention from the community. The ability to localize sound sources is a crucial aspect of auditory perception, it creates a sense of space for the listener, and provides information about the spatial location of objects in the environment. And more importantly, from a computational perspective, sound localization is the first stage of the auditory perception and undeniably represents one of the pillar of robot audition. The stage of sound localization generally conditions the good progress of the auditory interaction, since sound separation and speech processing rely or can be improved by the initial localization step. Thus, by strengthening and ensuring accurate sound localization abilities, the interaction process is greatly simplified, which increases the robustness of the aimed application, whether it be for interaction, navigation or scene mapping. Consequently, solving robot audition limitations, more or less comes back to solve the problem of sound source localization in realistic environments.

Chapter 2

Sound source localization for robotics

This chapter introduces the issues of sound localization in robotic context. As underlined by the previous chapter, one of the toughest challenge for machine hearing consists in finding from where the sound source originates. This task is particularly complex in realistic environments, inducing interference from reverberation and fluctuating acoustic conditions. The focus on this particular topic is easily explained by the fact that most of the interactions and applications based on hearing sense are conditioned by an accurate localization. Naturally, the studies of human auditory sense whether it be for localization, separation or recognition, as illustrated by the *cocktail party effect*, largely inspired robot audition.

The current chapter reviews, first, the human hearing ability in Section 2.1, through the anatomy of the auditory system and the physiological hypotheses that explain the hearing process. More specifically, this section focuses on the mechanism involved in the sound localization process through computational models and the related auditory cues.

In a second phase, we focus on how the knowledge of sound localization mechanism is transferred to robotic context. In detail, Section 2.2 introduces the low-level auditory cues and their processing for machine hearing and by extension for robot audition.

The last segment of this chapter expounded in Section 2.3 reviews the current progress and limitations of sound source localization in robotics. We evaluate the two fundamental paradigms of sound localization in robotics: on one hand, the binaural approach that is in general inspired from the physiological knowledge about mammals auditory systems, and on the other hand array processing approaches stemmed from signal processing and sensor fusion.

2.1 The human auditory system as source of inspiration

2.1.1 The human auditory system

Similarly to many aspects of human sensing skills, the human auditory system is a relevant illustration of ability for sound localization, and separation. However achieving this level of performance from a computational perspective, is still unreached from the current tools addressing robot audition. For the past centuries, the human auditory system has been a major topic of research and inspiration for psychoacousticians. Understanding the latter auditory capabilities is one of their pursued goal. An initial stage for building an artificial auditory system requires a thorough understanding of the biological structure of hearing organs and the neural activities associated to the auditory sensation.

As described in Chapter 1.2.1, the sound is a wave propagating in a medium. This wave when reaching the ears should be interpreted by the auditory system into an exploitable and understandable information. The sensation of sound is caused by varying pressures in the ear. Based on the airborne nature of the auditory information, the auditory perception is a range sensing, that played a crucial role during species evolution. Hearing organs for airborne sound probably appear around 250 millions years ago [SC09], during the Triassic era that corresponds to a period of transition from water to land. On the land, hearing organs evolved separately among species and mammals were endowed with tympanic ears that provided the more "recent" sensing skill. Hearing sense stimulated afterwards the rise of communication and languages. Until now, for many mammals species the sound localization skill is essential for detecting preys or conversely predators when vision is limited by a restrained field of view, obstacles or obscurity. Then, it is not surprising that our ancestors evolved sensitive tympanic ears and highly specialized auditory organs capable of processing airborne sounds.

2.1.1.1 Mechanical structure

The auditory system transforms the sound wave into neural activity. For this purpose, the auditory system is physically composed of three parts that are respectively the outer, the middle and the inner ear illustrated in Figure 2.1. The outer hear composed of the pinna and the ear canal principally amplifies the sound wave before transmission to the middle ear. The elastic cartilage that builds up the pinna is involved in the sound localization process with complex spectral filtering patterns governed by the shape of the pinna as stated in [Ray07] (see Section 2.1.2). The ear canal that has a resonating conic shape transmits the filtered sound to the tympanic cavity through the tympanic membrane. Due to the resonance in the auditory canal the pressure of the sound wave is increased before reaching the tympanic cavity [Wie47]. In the middle ear, the sound pressure is therefore transformed into vibration by the eardrum (*i.e.*, tympanic membrane) transmitted to the auditory ossicles. The auditory ossicles consist in three small bones that are the malleus, the incus and the stapes. As sound waves vibrate the eardrum, it in turn moves the nearest ossicle, the malleus, to which it is attached. The malleus then transmits vibrations,

via the incus, to the stapes, and ultimately to the membrane of the oval window, on the outside of the cochlea. The cochlea that builds up the inner ear is filled with

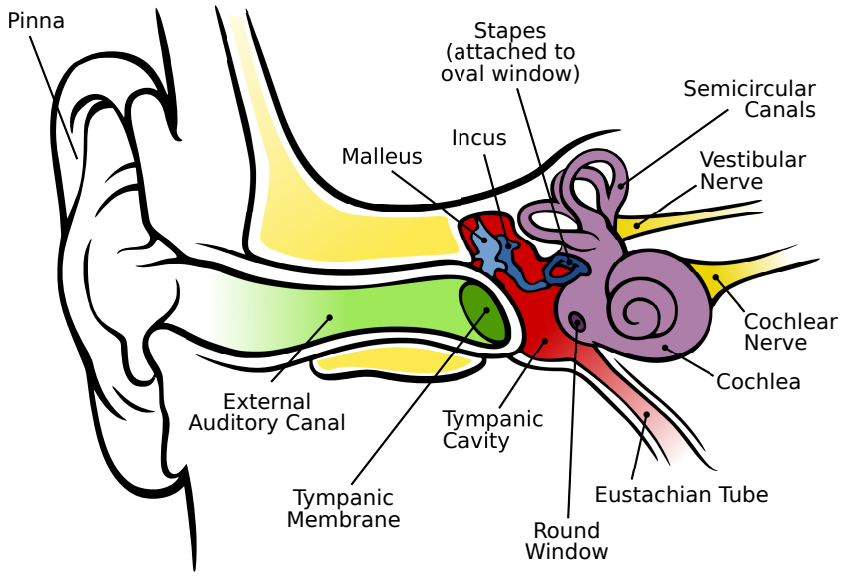
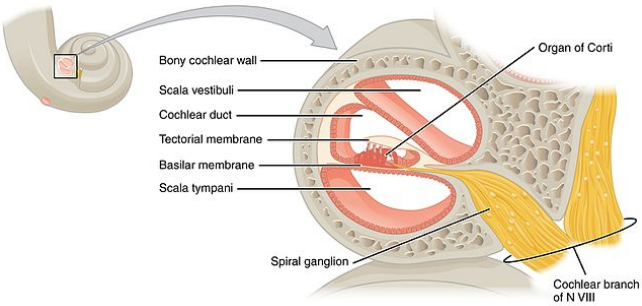


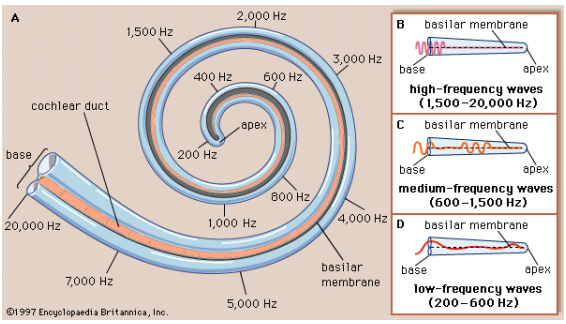
Figure 2.1: Anatomy of the human auditory system as represented in [CB05]

a salty watery liquid, the perilymph, which moves in response to vibrations coming from the middle ear via the oval window. The cochlea is sealed by the round window membrane, which vibrates with opposite phase to vibrations entering to the inner ear through the oval window. The mechanical amplification of sound waves through the ossicles is essential since it allows to improve the transmission of the sound wave from air to liquid medium. It is said that without ossicles, only 3% of the sound wave would be transmitted to the cochlea [HWA08]. It is not surprising that damaged or absent ossicles lead to a moderate-to-severe hearing loss [AGF+99]. Thus, the presence of the ossicles to concentrate the force of the vibrations improves the sensitivity to sound. The exact functioning of the cochlea is still a discussed subject in scientific communities. A first mechanical model of the cochlea has been proposed in 1863 by von Helmholtz, while the current mechanical model of the cochlea is stemmed from [VBW60] introduced one century later. The cochlea consists in a coiled tube enclosed in a hard bony shell split into two canals by the basilar membrane a illustrated in Figure 2.2. These two canals, respectively named vestibular and tympanic canals, compress and dictate the motion of the perilymph thanks to the alternate vibrations of the oval and round windows. The basilar membrane model is equivalent to a set of harmonic oscillators tuned to different frequencies from 20 Hz at the apex to 20 kHz [MOD89]. The basilar membrane resonates accordingly to the frequency of the signal thanks to different stiffness from its base to its apex that is more flexible. By combining this property to the inertial effects of the perilymph (*i.e.*, the frequency of the sound wave), that transmits a phase delay of a few cycles to the basilar membrane oscillation, the cochlea acts as a filterbank which transforms vibrations into

a neural spike train. The spike train has the same phase as the sound vibration [Wev49]. Inside the basilar membrane, the organ of Corti filters the sound. Indeed the motion of the basilar membrane stimulates the hair cells contained in the organ of Corti. The 15000 hair cells that are contained in the organ of Corti are the hearing receptor cells that are sometimes compared to photo-receptor cells of the eye retina. These hair cells are from to types: the outer cells play the role of amplification, while the inner cells transform the vibration into electric signals that are transmitted to auditory nerves. The information passed to the brain is represented as a cochleagram that is equivalent to a spectrogram (see Section 2.2.2).



(a)



(b)

Figure 2.2: The cochlea: In (a) transverse view of the cochlea [B+16] and (b) basilar membrane functioning (www.britannica.com)

Yet, the crucial point of the auditory system remains the interpretation of this neural activity. Actually, the neural activity travels from the cochlea to the auditory cortex where the signal is decoded and interpreted in an auditory sensation (localization, sound identification...).

2.1.1.2 Neuronal path

The brain has to solve the difficult task of transforming the acoustic wave input into meaningful auditory sensation. As already depicted in Figure 2.1 the sound signal travels from the ear to the auditory cortex. The auditory information in the brain is then processed in the midbrain, that is divided in several nuclei. The superior olivary complex (SOC) is the first station that receives input from both

ears through the connections with the cochleas. The SOC is the very early stage of signal interpretation since interaural comparisons are processed [TY02], which is an important source of information for the auditory system for determining the sound location. The SOC is divided into two nuclei that are respectively the lateral superior olive (LSO) and the medial superior olive (MSO). The MSO seems dedicated to process the time difference between the sound wave as exposed in [JSY98], while the LSO is more devoted to the difference of level observed between the sound wave [JY95]. As detailed in the next section, these two types of information are used for localization in the azimuth plane. This information coupled with the elevation information induced by the pinna is afterwards integrated to the Inferior colliculus (IC) where a 3D spatial representation of the source location can be inferred [CFC02]. The IC is also connected to the superior colliculus and transmits the localization information for the attention mechanism [ZVVO04].

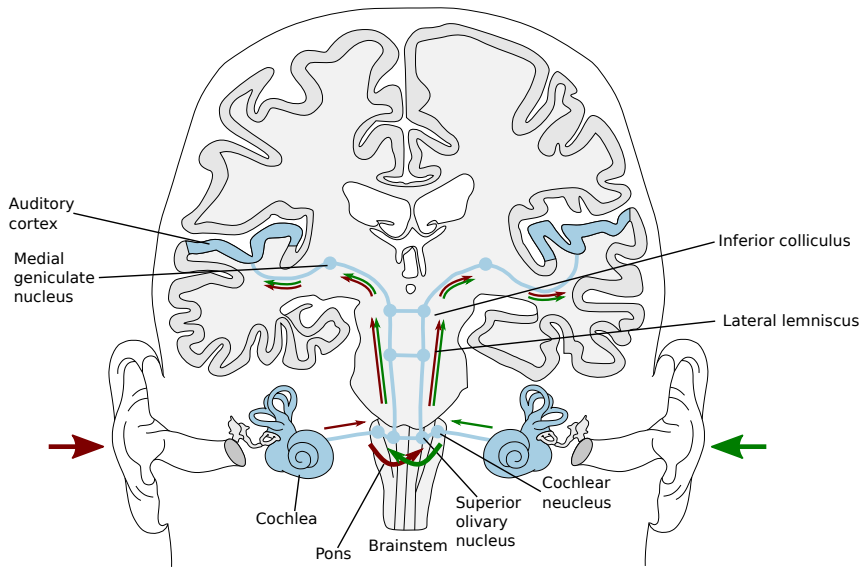


Figure 2.3: Neuronal path of sound from the ear to the auditory cortex. Figure adapted from http://twoears.aipa.tu-berlin.de

Eventually this information is transferred to the medial geniculate nucleus, before reaching the corresponding auditory cortex. In the auditory cortex, the spatial layout of frequencies in the cochlea along the basilar membrane is repeated through what it is called the tonotopic organization [TSM+04].

Globally this neural scheme stresses a central process of sound where the localization seems to be performed by a spectral interaural comparison. This interaural comparison is the core idea behind sound localization in robotics and in machine hearing in general.

2.1.2 Physiological mechanisms of sound localization

2.1.2.1 Binaural cues

Now, one can focus on the localization process that is performed in the SOC. In the case of sound source localization, the interaural cues have been identified as carriers of the spatial information of the perceived sound. These assumptions have been formalized through the *Duplex Theory* in [Ray07] under the assumption of a perfectly spherical head without any external ears (pinnae). This theory explains many properties of human sound localization. The identified interaural cues are respectively the interaural time difference (ITD), that is generated by the delay between the two ears when sensing the auditory event, and the interaural level difference (ILD) that is caused by the attenuation of the sound level between the two ears (see Figure 2.4). Both ILD and ITD are related to the sound source direction of arrival in the azimuth plane. If both ITD and ILD cues are available, the ITD cues prevail as shown in experiments conducted in [WK92] where conflicting cues where presented to listeners. ITDs and ILDs are complementary cues that are relevant in different ranges of

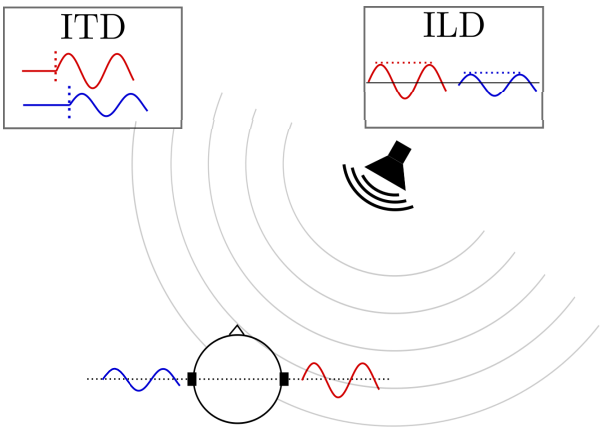


Figure 2.4: The ILD is caused by the difference of sound pressure between the two ears while the ITD reflects the phase difference between the signal perceived by the two ears

frequencies, as illustrated in Figure 2.5. Actually, ILDs are significant for frequencies above approximately 1.5 kHz, that is to say when the head is large compared to the wavelength λ of the incoming sound. In this case, the head shadows most of the incoming waves, by reflection mechanisms already discussed in the previous chapter. Consequently the perceived sound energy is drastically attenuated for the furthest microphone from the source. Conversely ITDs are meaningful for low frequencies. ITDs can be interpreted unambiguously only for frequencies for which the maximum physically-possible ITD is less than half of the wavelength λ, which corresponds somehow to the Nyquist frequency. Otherwise ITDs are subject to spatial aliasing related to the phase ambiguity. For a typical human head size (*i.e.* ≈ 22 cm between each ear) the maximum possible ITD is about 660μs, which sets the frequency limit of meaningful ITDs to around 1.5 kHz. However, the frequency ranges for which

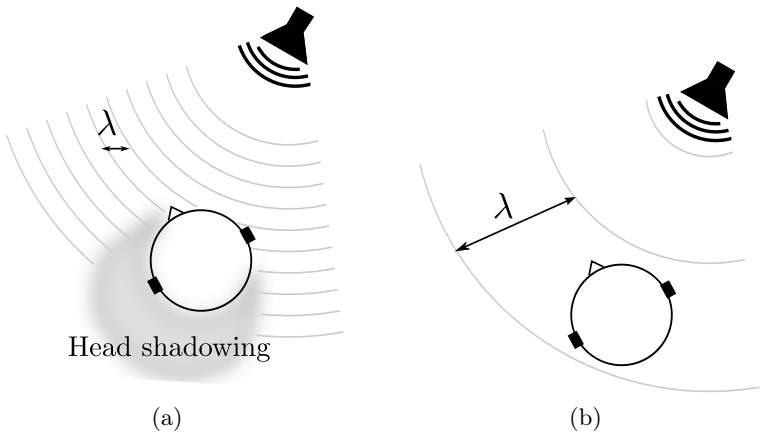


Figure 2.5: The ILD is consistent at high frequencies (a) since the shadowing effect of the head induces significant sound level difference. On the other hand ITD suffers from aliasing since the small wavelength λ induces phase ambiguity. At low frequencies (b), the aliasing is suppressed but the shadowing effect of the head is not significant.

the auditory system is sensitive to ITDs and ILDs significantly overlap, and most of useful sound signals are wideband (*i.e.*, include low and high frequency components). The auditory system in most cases combine information from both ITDs and ILDs to extract the location of a sound source [Bla97]. Humans are highly sensitive to small differences in ITDs and ILDs. For low-frequency pure tones, for example, the minimal noticeable difference for ITDs is on the order of 10μs while for ILDs this value is on the order of 1dB, as underlined by studies such as [DC77] and [HD69]. These differences corresponds to minimum audible angles from 1° to 3° approximately. These results emphasize the fact that such auditory system is particularly acute for extracting and interpreting the source azimuth direction from the binaural cues.

Several neuro-computational models have been proposed in the psychophysics literature to explain how the brain extract the interaural cues. The time-delay neural network of Jeffress architecture [Jef48] is still at the core of current modelling of how humans perform binaural localization using ITD cues. This theory characterizes the MSO neurons functioning as a coincidence mechanism. Specifically, a mechanism consists of a number of central neural units that record coincidences in neural firings from two peripheral auditory-nerve fibers, one from each ear, with the same frequency. Furthermore, neural signals from the two ears are delayed by given amounts fixed for each pair of fiber, similarly to the architecture depicted in Figure 2.6. Because of the frequency synchronization, a given binaural coincidence-output unit at a particular frequency is maximized when the ITD at that frequency is exactly compensated for by the internal delay of the fiber pair. As a result, the ITD of a sound stimulus can be determined from the delay that has the greatest output over a range of frequencies. The short-term average of a set of coincidence outputs at a given frequency with respect to their internal delay corresponds to some extent

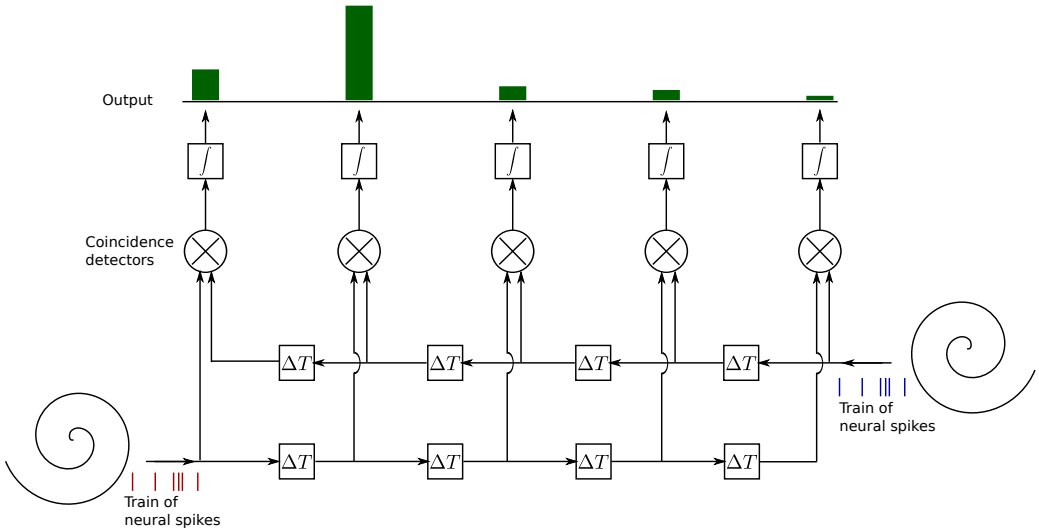


Figure 2.6: Jeffress neuronal architecture

to a short-time cross-correlation function of the neural signals arriving from the left and right ears. The utilization of such a mechanism has been validated for owls auditory system in [CK88]. It has been found that MSO neurons of the owls are connected in a very similar way to Jeffress diagram. The neurons communicating signals from the two ears converge in a orderly way, by increasing the delay in the transmission length. Furthermore, to account for the influence of ILDs in the latter model, Jeffress proposed the complete the neural mechanism with the latency hypothesis which is based the observation that the neural response to more intense sounds tends to start more rapidly than the response to less intense sounds, which implies that ILDs are somehow converted into ITDs. However, it has been suggested recently in [MG03] that mammals auditory systems do not fully comply with Jeffress model. The authors of the latter paper exhibit evidence of neural broad sensitivity to just two ranges of ITDs, that could subsequently describe a continuous space of possible ITD values. Likewise, a number of researchers have also reported cells that appear to respond to ILDs at several levels of the brainstem [BT68, CK83]. These researches suggest that ILDs could be detected by a unit which has an excitatory input from one ear and an inhibitory input from the other. Nonetheless the Jeffress model still remains acknowledged, and several auditory models are inspired from this pioneering work. This model stresses the cross-correlation mechanism between the two neural inputs that is underlay by the coincidence principle. In [Che61], the authors proposed a cross-correlation method directly based on the sound signal input rather than the neural input, in order to relate the interaural cross-correlation to the sound location. Since the cross-correlation is mainly designed to extract ITDs, they considered an additional mechanism to account for the effects of ILD, through a ratio of energy of signals measured on the left and right ear. Some other auditory models are oriented towards the neural response of the signal. This is the case of [Col73, Col77] in which the author characterizes the binaural interaction

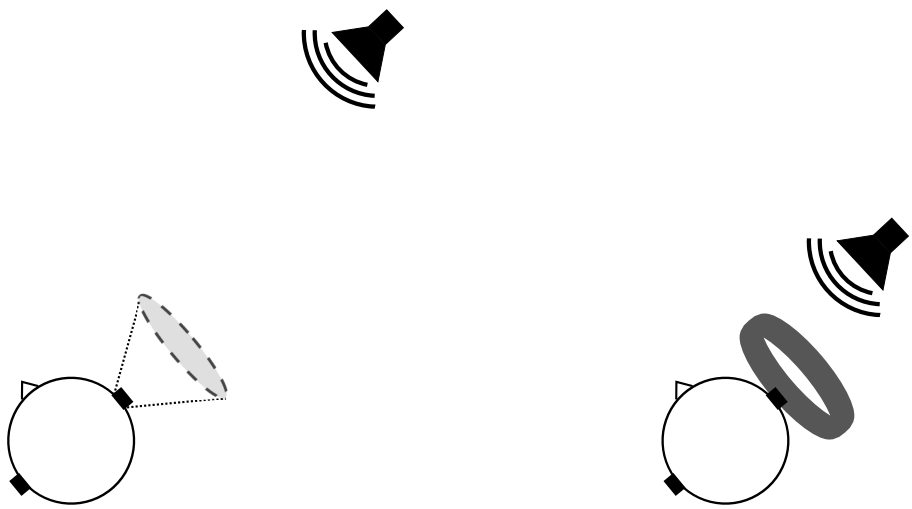


Figure 2.7: In the far field, localization based on interaural cues lead to a cone of confusion. In the near field the cone of confusion degenerates into a torus of confusion (black donuts) principally because of the action of the ILD (*i.e. intersection of the cone and a sphere*).

through a model of the auditory-nerve response to sound, and a comparison of the auditory-nerve response to the signals of the two ears. The latter model has also been extended in [SJC78] to include explicit assumptions concerning time-intensity interaction, a mechanism for extracting subjective lateral position, coupled with a frequency consistency mechanism. In the continuity of these works, the authors of [Bla80, Bla97, Bod93] proposed more complex binaural hearing models able to cope with real environments (*i.e.*, *cocktail party effect*)

Unfortunately, despite this variety of hearing models, even the most complex neuro-computational cannot fully explain the performance and accuracy of human auditory system. This fact strengthens the idea of interaction of several cues and mechanism for sophisticated localization process. This idea is confirmed by various experimental studies [WK89, SN36, WAKW93] that stress significant errors of localization (in static configuration) when a listener is asked to indicate if the source stimuli is produced in the front or in the back. This is due the so-called *front-back ambiguity* that is inherent to the binaural cues. The *front-back ambiguity* may be considered as the main limitation of the interaural cues. This ambiguity implicitly confirms that there exists multiple source locations that share the same ITD and ILD cues.

As expounded in [vHW20, Bla97], at it simplest form, iso-ITD locations tailor a hyperbolic surface of rotation symmetric to the interaural axis. In the far-field (*i.e.*, distances greater than 1 m for a typical head), the hyperbolic surface approximates a cone centered on the interaural axis, that shapes the well-known *cone of confusion*. Empirical measurements of the ITD with respect to the source direction also confirm the cone-shaped iso-ITD [Mil72]. On the other hand, in the near-field low frequency ILDs are constant on spheres centered on the interaural axis [SCSK00], while these

ILDs become perceptually insignificant as the source gets farther from the head. Consequently, positions that give rise to the same binaural cues (i.e., constant ITD and ILD contours) in the near-field form a torus that has been described in [SCSK00] as *torus of confusion*. The *torus of confusion* degenerates to the *cone of confusion* for sources in the far-field as the influence of low-frequency ILD is diminished. This confusion can be solved by the motion of the head [Wal40] (see Section 2.1.2.3)

2.1.2.2 Monaural cues

As a consequence of the previous results, it can be inferred that interaural cues do not fully described the localization process. Some complementary cues are necessary to disambiguate the source position in the cone of confusion. Actually the localization process is also described by other dimensions that are the elevation and the distance to the source. Several circumstantial evidence tends to relate the sound stimuli localization to an action of the pinna. As an illustration, experimental results exhibited in [OP86] showed that localization performance is substantially degraded when the localization is performed with a single ear. More specifically, most of the azimuth localization were erroneous but elevation localization remained consistent, showing that some localization ability is somewhat preserved in the monaural case. After some training the azimuth could be recovered (however without the same accuracy as with a binaural auditory system). This typical experiment emphasizes mechanisms of localization implied by monaural cues, that principally concerns elevation.

The following attempt [Bat67] aiming to characterize the effect of pinna filtering in localization process, revealed that sounds reaching the ears interact with the physical characteristic of the listener, that is to say, head, shoulders and upper torso. The modifications induced by these interactions can be subsequently used to estimate the direction and distance of the sound source. All together, these interactions characterize a complex response function known as the head-related transfer function (HRTF). Analogously, in the temporal domain, the HRTF corresponds to the head-related impulse response (HRIR) (see Figure 2.8). HRTFs have been proven to influence sound localization cues (*i.e.*, ITDs, ILDs) and the spectral shape of the sound reaching a listener [Har99]. Indeed the sound needs to propagate around the head (diffraction) the reach the ears. HRTFs establish the relationships between the sound in the free-field zone and the sound perceived by the ears of a listener [WK89, Bla97]. For instance in [WK89] the authors used probe microphones in the ear to measure and describe the transfer function from sound source to eardrum in anechoic environments. Hence HRTFs vary significantly between different individuals due to differences in anatomy. Such influence characterizes the individual variability of sound perception as stated in [Bla97]. An HRTF itself can be decomposed into two separate components: the common transfer function (CTF), which is common to all sound source locations [MG90] (*i.e.* depending on the anatomy of the listener) and the directional transfer function (DTF), which is related to the sound source direction. Explicitly, when a sound source is situated in the near-field, the HRTF is simultaneously depending on the direction and the distance of the source across all frequencies [BR99]. In the far-field the dependence of HRTFs on distance vanishes.

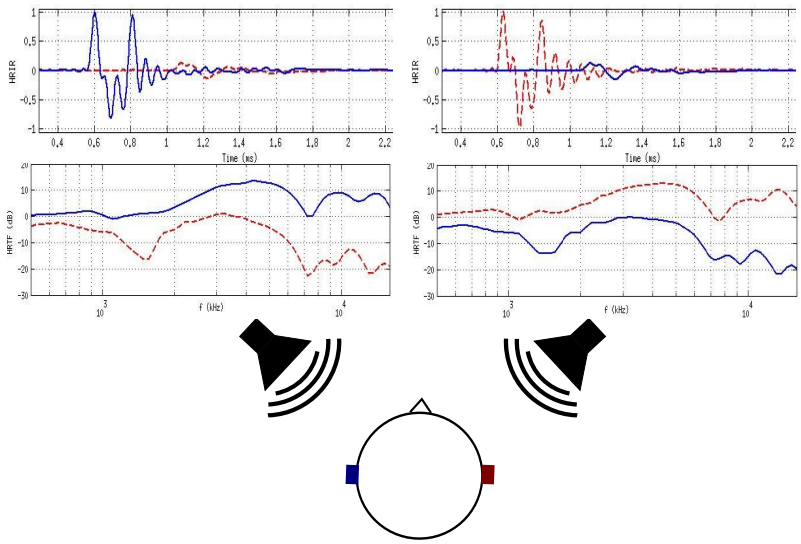


Figure 2.8: The HRIR (top), as well as the HRTF (bottom) captures the relationships between a measure of the signal in the free space and the signal perceived by the ear. The HRTF (respectively HRIR) depends on the source location and frequency. Data plotted from the Cipic database [ADTA01].

As a result, the HRTFs of each ear are then functions of four variables: the frequency of the sound source, the sound source azimuth and elevation angles, and the distance to the sound source as stated in [ZDD04].

With the use of HRTFs, many of the localization limitations (*e.g., cones of confusion*) inherent to models based on the use of ITD and ILD alone are overcome. Although elevation measurements mechanism has received less attention than localization in the horizontal plane (*i.e.*, through interaural cues), the hypothesis that the elevation is recovered through the spectral filtering action of the pinna is now acknowledged [MOD89, PB93]. This filtering is especially useful to determine the localization of the source in the vertical plane, and the resolution of *front-back ambiguities*. In the first case, there are substantial psychoacoustical evidence to support the hypothesis that pinna spectral notches are important cues for determining the elevation of the source [MOD89, WHW74, TY03]. The sharp notches in the signal spectrum, commonly referred to as pinna spectral notches, are related to the elevation of the sound source [RDY05]. In [CP94] it has been shown that the locations of spectral notches near 7 kHz change as a function of elevation, as illustrated in Figure 2.9. Moreover, the asymmetry of the pinna could help to distinguish source from the back or front side. On the basis of the reflections of the various ridges of pinnae, different filtering effects are observed depending on the side on which is located the sound source [MG91]. The selective sound shadow of the pinna in resolving the *front-back ambiguity* problem was confirmed in [FF68, PN97b]. However, because of the high variability of size and shape of pinnae between listeners, the task of relating the anatomy of pinnae to localization cues they create remains elusive. Hence, it is impossible to infer the sound location (azimuths and elevations) from unknown

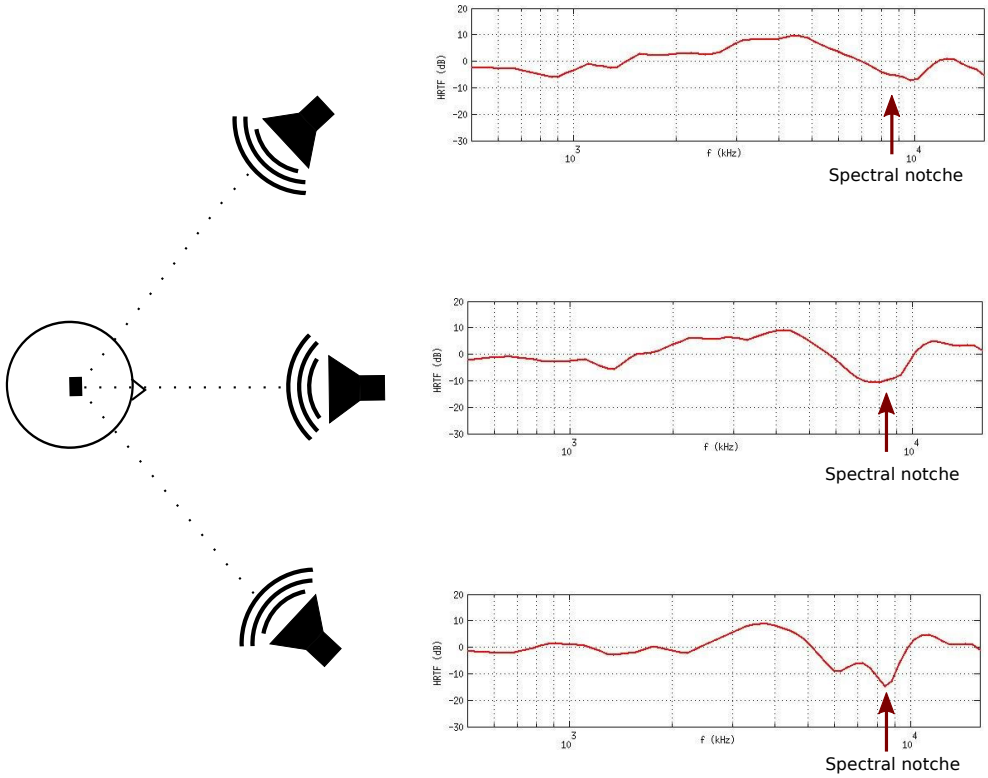


Figure 2.9: The spectral notches in the HRTF, are related to the elevation of the sound source. Data plotted from the Cipic database [ADTA01]

HRTFs.

A further complication associated with judging the location of a sound source in enclosed room, arises from sound reflections from various surfaces before reaching the ears of the listener. However, despite the fact that reflections originate from many directions, listeners are able to determine the location of the direct sound quite accurately. The auditory system is thus able to differentiate competing sources, the direct sound and its reflections, for localization. This ability during the localization process is referred to the *precedence effect* [WNR49] that characterizes the masking of echoes direction in the localization process. Early reflections are spectrally and temporally similar to their source signals but however carry spatial acoustic cues unrelated to the source location. Despite this correlation, a single reflection arriving within 5 to 30 ms can be up to 10 dB louder than the direct sound without being perceived as a secondary source [Yos94]. Hence, localization cues are mainly based on the first sound wave that is given a higher perceptual weighting than the reflections.

Eventually distance estimation is processed from monaural cues. Naturally, the sound intensity is used for estimating the distance, particularly when the sound stimulus originates from a well-known source. Indeed as already mention in the previous chapter, the typical sound decay certainly provides the most intuitive cue to estimate distances. Several studies [War68, SG62] showed experimentally the

correlation between the perceived distance and the loudness of the source, that is perceptually related to the intensity. In enclosed rooms, the distance to a sound source can also be estimated from reverberation. The ratio of the direct sound over the first reflections, also known as Direct-to-reverberant ratio (DRR) or the time gap between the direct sound and the first reflections are assumed to be cues for distance estimation [She82]. The effect of the reverberation has also be denoted in [Nie92] where distance judgments where much more veracious under reverberant conditions than anechoic conditions. We refer the interested reader to a more exhaustive review provided in [ZBB05] addressing distance cues.

2.1.2.3 Influence of the motion in the localization process

The previous chapter emphasized the constraints related to motion of robot, that requires to be able to process the sound dynamically and in real-time. However a question rises to know if the motion is an actual constraint or conversely an advantage to the localization process. Thus in the following, we study the evidence related to the motion influence for human localization process. The individual variability of the auditory cues [Bla97] reveals that they do not explain solely the performance of the human auditory system. More importantly, it suggests that other mechanisms are involved in the perception of the sound source. Thus some other modalities must be taken into consideration. Among them, the self-motion plays an important role for the correct sensing of the sound event. Indeed, in daily sound listening, our heads are constantly moving, which tends to emphasize the potential influence of the motion.

From this perspective, almost all the studies agreed on the fact that localization through motion performs better than localization in a static way. In [Wal40], it has been shown that the accuracy of sound localization is improved if listeners are allowed to move their heads during the sound stimuli. A prior work [Wal39] also demonstrated that for longer duration sounds, head motion cues dominate other cues. Some other studies like [TMR67] suggests that the accuracy of the sound localization is improved by the head motion. Compared to static cases, they report an improvement of 10 to 15%, especially when the sound stimulus is long enough or when the source is located in eccentric positions.

A first thought could assume that the motion multiplies the auditory cues measurements and build up a spatial representation of the source location by fusing the measurements. As a consequence the localization should be more robust by using the redundancy of the measurements. However it appears from different experimental studies that the dynamics of the auditory cues, more particularly the dynamics of the interaural cues, are also a keypoint of the localization performance.

Actually, the motion contributes to the disambiguation of the *front-back confusion* [WK99] inherent to the interaural cues. Indeed the variation of ILDs or ITDs with respect to the head motion, characterizes the fact that the sound source is in front or in the back of the listener. Moreover, the dynamics of the interaural cues also plays a role in the estimation of the elevation of the sound source. When the head is rotated, ITDs and ILDs change differently depending on the sound elevation [PN97a]. More specifically, the latter study showed that localization within *cones of confusion*

remained possible despite distorted HRTFs, thanks to the head movements. Likewise, head motions (translations) induces change in azimuth that is range dependent. For sources that are very close, a small shift causes a large change in azimuth, while for sources that are far the azimuth variation is minimal [ADN95]. In the same way, [LHC90] suggests that the distance can be perceived dynamically through motion. Actually, when the level (loudness) of the sound source varies with varying distance, the apparent distance judgments of the listener change accordingly. In an anechoic environment, this relationship is characterized by the inverse-square law: the level of the sound falls by more or less 6 dB for each doubling of the source distance [Col63]. There are also many other ways in which motion might help in audition: moving towards a source can increase the direct to reverberant ratio (*i.e.*, increase the speech intelligibility as stated in Chapter 1.2.2.2) while head rotation can position the target source in the frontal plane where spatial resolution is at the greatest [CLH97] compared to peripheral locations. This latter property is often relate to the *auditory fovea*, by reference to *visual fovea* at the center of the retina that is capable of higher resolution that in the periphery.

These studies emphasize the strong link between the motion and the auditory cues. Similarly, a study of [VVGVO04] involving a moving sound source, showed that acoustic inputs are continuously combined with feedback signals about changes in head orientation. The motion might not be pre-programmed as it could be assumed but rather dynamically adjusted by the changes of the auditory cues. These results support the theory of a real-time feedback loop between the motion and the auditory cues. Moreover, the dynamics of the interaural cues seems to be less sensitive to listener's HRTFs as exposed by [KUKH03]. The authors suggests that the information induced by the head motion is useful for overcoming changes in listener's HRTFs. Globally, the auditory auditory scene analysis appears to be a dynamic process that is influenced by the active sensing of the environment as stated by [KPTK12]. The auditory process can then be considered as a sensorimotor loop where the motion and the auditory perception are connected.

2.2 Localization cues for robot audition

2.2.1 Modelling the spectral information

The main cues use for machine audition are derived from the knowledge of the mammals auditory system. It is not surprising that most contributions in robot audition context are based on interaural cues. A first step before extracting the interaural cues concerns the representation of the sound signal recorded by the microphones. The signal can be naturally exploited in the temporal domain, but for more efficiency and flexibility, the spectral decomposition of the signal is preferred. This way to process the sound draws closer to auditory mechanisms in the ear and cochlea. The simplest way to transform the temporal signal spectrally is by a Fourier transform. However, since sound signals used in robot audition are generally broadband and non-stationary, the short-time fourier transform (STFT) provides more relevant information. Simply, the function to be transformed is multiplied by a window func-

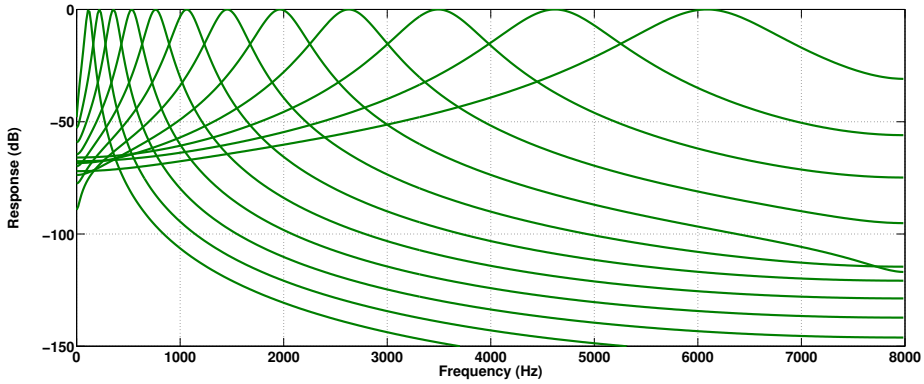


Figure 2.10: Gammatone filters frequency response based on Lyon's cochlear model [Ell09]

tion which is nonzero for only a short period of time. The Fourier transform of the resulting signal is taken as the window is slid along the time axis, resulting in a two-dimensional representation of the signal. Thus, let us consider the signal $x_1(t)$ and $x_2(t)$ respectively measured by two microphones $\mathbf{M}_1$ and $\mathbf{M}_2$. In this case the STFT $X_i(f,T)$ at a given frequency f of the signal $x_i(t)$ is simply defined as

$$X_i(f,T) = \int_{-\infty}^{+\infty} x_i(t)w(t-T)e^{-2j\pi ft}\mathrm{d}t, \tag{2.1}$$

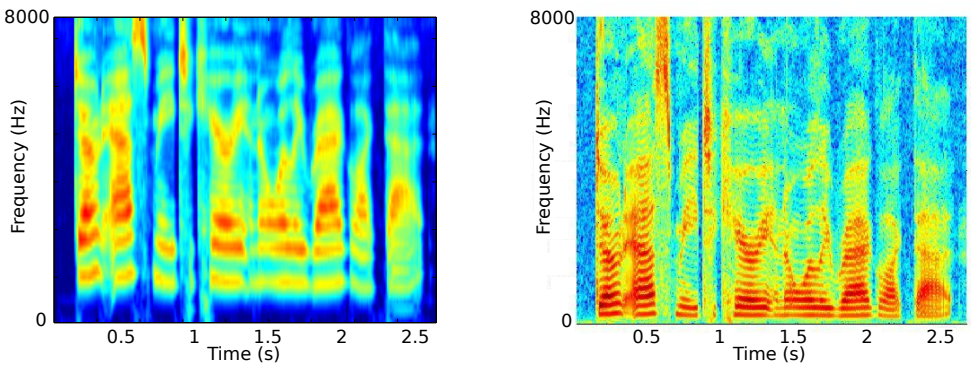
in which $w(t-T)$ stands for a window function (*e.g.*, Hamming, Hann, sinusoidal...) centered on T sliding along time axis. For discrete-time signals as generally processed in machines, the STFT consists in computing the Fourier transform on shorter segment of a longer signal, divided in equal length segments beforehand. Assuming that $t = 1,\ldots,T$ and $f = 1,\ldots,F$ are, respectively, time frame and frequency bin indices. With these notations, (2.1) becomes

$$X_i(f,l) = \sum_{t=0}^{N-1} x_i(t)w(l-t)e^{-2j\pi ft/N}, \tag{2.2}$$

where N is the window length, and $l = 1,\ldots,L$ characterizes the frame that is processed[1]. The STFT transforms the sound signal into a spectrogram, that represents all spectral components of measured sound (see Figure 2.11b). Nevertheless, the methods based on STFT do not perform a frequency decomposition similarly to the human inner ear. Actually as mentioned in the previous section, the components of the outer and middle ear, the signal is broken up into different frequencies that are naturally selected by the cochlea and hair cells. A frequency decomposition trying to mimic the effect of the basilar membrane inside the cochlea can be obtained from a gammatone filterbank [Joh72]. In this case, each filter is associated with a specific area on the basilar membrane. The gammatone filter is nowadays widely used

[1]For simplicity we use the notation $x(t)$ instead of the $x[t]$ for discrete signals

to model the cochlea filtering action as illustrated in Figure 2.11. The gammatone filter provides an impulse response as the product of a gamma function and a tone. The output of the filter can be regarded as a measurement of the basilar membrane displacement. Gammatone filters constitute a good approximation of human spectral analysis at moderate sound levels [PAG95]. Their bandwidth increases with frequency (see Figure 2.10), mimicking human auditory ability to discriminate low-frequencies signals [SVN37]. It has also been shown in [PRH+92] that the impulse response of the gammatone filter fits to the human auditory filter shapes, derived in [PM86].



(a) (b)

Figure 2.11: Two different representation of the spectral components of a sound: in (a) the cochleagram based on gammatone filter banks and in (b) the spectrogram obtained by the STFT.

Mathematically, a gammatone filter is defined by its impulse response given by

$$g(t) = At^{n-1}e^{-2\pi Bt}\cos 2\pi ft + \phi \tag{2.3}$$

where A is the amplitude, n the order of the filter and B is the bandwidth of the filter that determines the duration of the impulse response. The filter bandwidth B is usually set according to the equivalent rectangular bandwidth (ERB) given by:

$$ERB(f) = \left[\left(\frac{f}{Q}\right)^p + (B_{min})^p\right]^{\frac{1}{p}}, \tag{2.4}$$

in which Q is the asymptotic filter quality at high frequencies and B_{min} is the minimum bandwidth for low frequency channels. The parameters of the ERB are generally adjusted from human auditory data. Hence in the literature, there exists several modelling such as the Glasberg and Moore cochlea model summarized in [GM90] ($Q = 9.26$, $B_{min} = 24.7$, $p = 1$), the Lyon's cochlea model given in [Sla88] ($Q = 8$, $B_{min} = 125$, $p = 2$) or the Greenwood model [Gre90] ($Q = 7.23$, $B_{min} = 22.85$, $p = 1$). Eventually the bandwidth of the filter in (2.3) is approximated as

$B = 1.019ERB(f)$. Hence the discrete-time cochleagram is obtained as

$$C_i(f,l) = \sum_{t=1}^{t=N} x_{ci}(t,f)w(l-t) \tag{2.5}$$

with $x_{ci}(t,f)$ defined as the gammatone-filtered signal. With this modelling, the cochleagram can provide better results when extracting auditory cues as demonstrated in [SM15, YAZ12b].

2.2.2 HRTFs modelling

Similarly to human listeners, the HRTFs also characterize robotic platforms, that can shadow, reflect or refract the sound, which modifies the sound perception. The modelling of the HRTFs is a big concern in robot audition, because of the modification of the interaural cues in a non linear way. The induced scattering and the spectral modification of the sound particularly affect the localization performance by altering the binaural cues. Indeed the filtering effect of the robot structure is reflected in sound level attenuation and in sound delays as explained in Chapter 2.3. This effect is particularly stressed by anthropomorphic robots that are endowed with a head, torso and pinna mimicking the human morphology.

Figure 2.12: Typical material required for HRTF measurements [ZX14]. In this solution the measurements are performed from two rings of speakers that rotate accordingly to azimuth and elevation angles.

Hence, the signals $x_1(t)$ and $x_2(t)$ measured by the microphones $\mathbf{M}_1$ and $\mathbf{M}_2$ are defined spectrally as

$$\begin{cases} X_1(f) = H_1(f)S(f) \\ X_2(f) = H_2(f)S(f), \end{cases} \tag{2.6}$$

where $X_i(f)$ corresponds to the Fourier transform of the measured signal, $H_i(f)$ the HRTF and $S(f)$ the Fourier transform of the signal emitted $s(t)$. However, it should be specified that HRTFs only account for the action of the robot structure (or body), on the sound, in free-field. Hence, HRTFs do not model acoustic effect such as reverberation. In reverberant conditions, (2.6) becomes more exactly

$$\begin{cases} X_1(f) = H_1(f)H_{1_{room}}(f)S(f) \\ X_2(f) = H_2(f)H_{2_{room}}(f)S(f), \end{cases} \tag{2.7}$$

where $H_{i_{room}}(f)$ characterizes the room acoustic effect such as reverberation. In this case $H_i(f)H_{i_{room}}(f)$ is the acoustic transfer function (TF) characterizing the difference between the emitted sound and the recorded sound.

HRTFs can be obtained from different ways. The straightforward way consists in building HRTFs from direct acoustic measurements of the signal at different location and frequency. The typical HRTF identification procedure consists in placing an artificial humanoid head and/or torso on a motorized turntable which can be rotated accurately to any required azimuth. A speaker mounted on a boom stand enables accurate positioning of the speaker to any elevation with respect to the humanoid. This solution is, however, not so convenient, since it requires anechoic conditions, and specific and generally expensive equipment to covers all the possible location of the sound source (see Figure 2.12). Furthermore the procedure requires long acquisition time. Fortunately, in the literature, several HRTF databases are available such as the CIPIC [ADTA01] and [WGS11] related to the dummy head *KEMAR*, or [TAM+02]. This solution provides accurate HRTFs but requires the same robot platform, with the same structure (*e.g.* presence of torso, pinna...) as the one used for creating the database, because of the high variability of HRTFs.

Nonetheless the exact HRTFs of the robots are not always accessible, due to their complex structure or the tedious procedure of measurements. As an alternative to the measurement process, a second solution consists in accurately modelling the body and head effect. When the robot has a very simple shape, HRTFs can be expressed through simple models such as the spherical model proposed in [Bla97]. For more realistic models, acoustic simulation software might used to estimate HRTFs through signal processing methods [OI06]. In robot audition, simplified models of the head have been proposed in the literature.

The auditory epipolar geometry (AEG) model [NOK00b] is the simplest one and considers free-field microphones in planar scene. In this model the interaural cues are directly related to the azimuth direction of the sound source since HRTFs are equal to

$$\begin{cases} H_1(f,\alpha) = 1 \\ H_2(f,\alpha) = e^{-2j\pi f \frac{d}{c}\cos\alpha} \end{cases} \tag{2.8}$$

in which d is the inter-microphones distance, c the sound celerity and α the azimuth of the sound assumed to be in the far-field. With $\mathbf{M}_1$ as a arbitrary reference, the auditory epipolar geometry states that signal measured by $\mathbf{M}_2$, $x_2(t+\tau)$ is delayed by $\tau = \frac{d}{c}\cos\alpha$, with respect to the reference signal $x_1(t)$. This model is by far the most used in sound source localization because of its simplicity, and the low computational

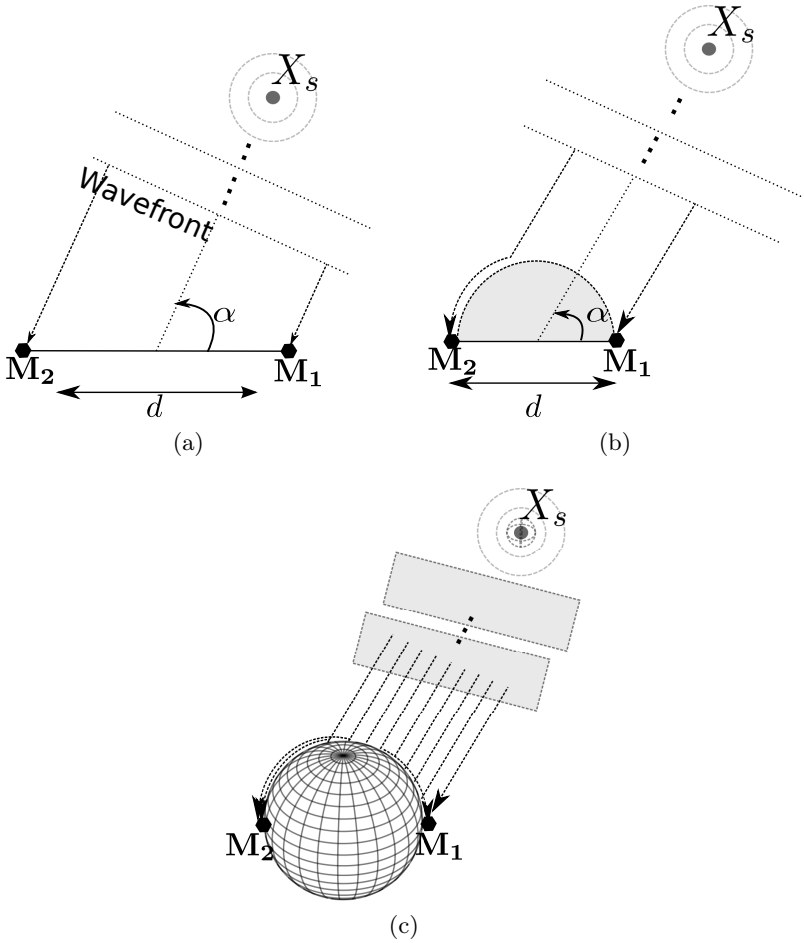


Figure 2.13: Head models characterizing the HRTF of the head: In (a) the auditory epipolar geometry considering free-field microphones, in (b) the revised auditory epipolar geometry considering a disk between the microphones and in (c) the scattering theory modelling a sphere between the microphones.

load that it requires for inferring the sound location. However this model assumes free-field installations and is generally used for array-based localization (see Section 2.3).

The revised auditory epipolar (RAEG) [NOK01] is an extension of the AEG by considering, this time, a disk between the microphones. The disk models the diffraction induced by the robot head before reaching the microphones. In this case, HRTFs are characterized by the following set of equations:

$$\begin{cases} H_1(f,\alpha) = 1 \\ H_2(f,\alpha) = e^{-2j\pi f \frac{F(\alpha)}{c}} \end{cases} \tag{2.9}$$

in which the shadowing effect of the head is modelled by the function:

$$F(\alpha) = \begin{cases} (\alpha - \frac{\pi}{2})\frac{d}{2} + \frac{d}{2}\cos\alpha \text{ when } 0 \leq \alpha \leq \frac{\pi}{2} \\ (\alpha - \frac{\pi}{2})\frac{d}{2} - \frac{d}{2}\cos\alpha \text{ when } \frac{\pi}{2} \leq \alpha \leq \pi \end{cases} \tag{2.10}$$

Through (2.10), $F()$ characterizes the path of the sound that follows the circular shape of the head (*i.e.*, diffraction) in order to reach the microphones as illustrate in Figure 2.13.

Eventually, the scattering theory (ST) proposed in [NMOK03], models HRTFs through a sphere diffracting the sound. Hence, the considered scene is not planar unlike the two aforementioned HRTF models. The scattering theory is based on the expression of diffraction of sound when considering an ideal rigid spherical head introduced in [DM98] as

$$\begin{cases} H_1(\ell, \beta, f) = \dfrac{\ell c e^{-2j\pi f/c}}{2j\pi f(\frac{d}{2})^2} \displaystyle\sum_{m=0}^{\infty} (2m+1) P_m(\cos(-\frac{\pi}{2} - \alpha)) \dfrac{h_m(2\pi\ell f/c)}{h'_m(d\pi f/c)} \\ H_2(\ell, \beta, f) = \dfrac{\ell c e^{-2j\pi f/c}}{2j\pi f(\frac{d}{2})^2} \displaystyle\sum_{m=0}^{\infty} (2m+1) P_m(\cos(\frac{\pi}{2} - \alpha)) \dfrac{h_m(2\pi\ell f/c)}{h'_m(d\pi f/c)} \end{cases} \tag{2.11}$$

in which ℓ characterizes the distance between the sound source and the center of the sphere. Furthermore the function $P_m()$ and $h_m()$ are respectively the Legendre polynomial of degree m, and the spherical Hankel functions of order m. $h'_m()$ simply denotes the derivative of $h_m()$.

These different models underline the additional complexity induced by HRTFs, that should be accurately modelled as well as the environment acoustic, to obtain exploitable auditory cues for humanoid robots.

2.2.3 Auditory cues

The ratio of HRTFs encodes binaural information. Indeed HRTFs depend on the direction of the sound that is implicitly related to the interaural cues. This relationship is not trivial when considering the actual HRTFs of robots. A typical solution consists in building a look-up table relating the ratio of HRTFs and direction of the sound source to interaural cues as exploited in [HLSVL06]. The relationship between the interaural cues and the sound location can also be obtained through the simplified HRTF models depicted above. In both AEG and RAEG, the ITD appears as a function of the azimuth angle of the source. The phase difference between the HRTFs is directly related to the ITD τ that can be extracted as follows in the AEG configuration

$$\tau_{AEG} = \frac{d}{c}\cos\alpha \tag{2.12}$$

when in the REAG, the ITD is defined as

$$\tau_{AEG} = \frac{d}{2c}\left((\alpha - \frac{\pi}{2}) + |\cos\alpha|\right) \tag{2.13}$$

However, these two models are centered on modelling the direct path in AEG and delayed path in RAEG of the sound source, that is to say the effect of the head model on the ITD. Consequently they do not provide any meaningful ILD. On the other hand, the ST strategy is more realistic by accounting for the delay and attenuation induced by a spherical head. Hence from this model both ITD τ and ILD ρ can be inferred from (2.9) as

$$\begin{cases} \tau_{ST} = \dfrac{\arg(H_1(\ell,\alpha,f)) - \arg(H_2(\ell,\alpha,f))}{2\pi f} \\ \rho_{ST}(dB) = 20\log_{10}\dfrac{|(H_1(\ell,\beta,f))|}{|(H_2(\ell,\alpha,f))|} \end{cases} \tag{2.14}$$

From the ST approach, more reliable azimuth directions can be extracted from ITD and ILD values. This reliability will strongly depend on the capacity to cope with environment acoustic (*e.g.* reverberation, noise..), that alters the latter auditory cues.

Prior to inferring the sound location from HRTF models, ITDs and ILDs should be estimated from the sound signal recorded. The coincidence mechanism of Jeffress auditory model inspired the techniques of interaural cues extractions that are commonly based on the cross-correlation of the two recorded signals. In the basic form, the interaural cues can be obtained directly from the temporal signals. Considering discrete sampling of the signal, the ITD is obtained from the standard cross-correlation between the signal $x_1(t)$ and $x_2(t)$ respectively recorded by the microphones $\mathbf{M}_1$ and $\mathbf{M}_2$ defined as

$$\boldsymbol{r}_{1,2}(\tau) = \sum_{t=1}^{T} x_1(t)x_2(t-\tau) \tag{2.15}$$

Hence the estimated ITD $\widehat{\tau}$ is obtained from the peak(s) of the cross-correlation function:

$$\widehat{\tau} = \operatorname{argmax} \boldsymbol{r}_{1,2}(\tau) \tag{2.16}$$

As ILD is concerned, it is simply defined as

$$\rho = \frac{\sum_{t=1}^{T} x_1^2(t)}{\sum_{t=1}^{T} x_2^2(t)} \tag{2.17}$$

that can be also expressed in dB as $\rho_{\text{dB}} = 20\log_{10}(\rho)$.

As outlined by the localization mechanism in mammals auditory system, ITDs and ILDs are inferred from the spectral representation of the signal. In this context, ITDs are generally estimated from the spectral domain with the Generalized cross-correlation (GCC) [KC76]. The Cross-Correlation function exposed in (2.15), becomes

$$\boldsymbol{R}_{1,2}(\tau) = \sum_{f}^{F} \psi(f)\phi_{x_1,x_2}(f)e^{\varphi(\tau)}. \tag{2.18}$$

where φ corresponds to the phase shift for a defined ITD τ. The cross-spectral power density ϕ_{x_1,x_2} is usually defined by

$$\phi_{x_1,x_2}(f) = \frac{1}{L}\sum_{l=1}^{L} X_1(f,l)X_2^*(f,l) \tag{2.19}$$

$X_1(f,l)$ and $X_2^*(f,l)$ are respectively the STFT of $x_1(t)$ and the conjugate of the STFT of $x_2(t)$. There exists also other alternative that defines the cross-spectral power density with a "max" pooling function so that

$$\phi_{x_1,x_2}^{max}(f) = \max_{l} \;\; X_1(f,l)X_2^*(f,l). \tag{2.20}$$

This solution can perform better for intermittent sound sources since it does not integrate irrelevant information when the source is inactive [BOV12], as it would be the case with (2.19). Eventually $\psi(f)$ is a weighting function, that generally aims to enhance the cross-correlation function. It should also be noted that when $\psi(f) = 1$, (2.18) simply corresponds to the standard cross-correlation. In (2.18), the maximum peak of the cross-correlation function gives an estimation of the actual ITD and can therefore be written as:

$$\widehat{\tau} = \underset{\tau}{\operatorname{argmax}} \; \boldsymbol{R}_{1,2}(\tau) \tag{2.21}$$

When k sound sources are active, the latter function returns several peaks wherein the k first peaks should correspond to the ITD of each sound source.

Depending on the signal of interest, there exists in the literature various weighting functions $\psi(f)$ that aim to enhance the sharpness of the peaks given by the cross-correlation. The most common weighting function used in robot audition is certainly the PHAse Transform (PHAT). The PHAT filter is a normalization factor that aims to increase the sharpness of the peaks. The filtering method is also known to be robust towards reverberation as demonstrated in several studies [Her86, DHD07] by preserving the information related to the phase. Mathematically, the filter is thus defined as follow

$$\psi_{PHAT}(f) = \frac{1}{|\phi_{x_1,x_2}(f)|}. \tag{2.22}$$

But the PHAT weighting is uniformly applied to all frequencies, without differentiating signal frequencies to noise frequencies. As a result GCC techniques using PHAT weighting functions are particularly sensitive to noise. Recently, extensions of the method have been proposed in order to control the degree of whitening and limit the amount of degradation from the independent noise with a parametric modification. One can cite $PHAT_\beta = |\phi_{x_1,x_2}|^{-\beta}$ where $0 < \beta < 1$ is the tuning parameter that control the level of weighting [MP10]. Similarly the Reliability-Weighted Phase Transform (RWPHAT) integrates a noise weighting function, a Wiener filter gain, based on the SNR of the signal [VMR06] (see Chapter 1.2.3.2).

Another classic weighting function is the ROTH processor [Rot71]. The weighting function is expressed as follows.

$$\psi_{ROTH}(f) = \frac{1}{\phi_{x_1,x_1}(f)} \tag{2.23}$$

The Roth processor has the property of suppressing the frequency regions where noise is large [KC76].

There is also the Smoothed Coherent Tranform (SCoT) filter given by

$$\psi_{SCoT}(f) = \frac{1}{\sqrt{\phi_{x_1,x_1}(f)\phi_{x_2,x_2}(f)}} \tag{2.24}$$

The SCoT method can be considered as pre-whitening filters followed by a cross-correlation. It should also be noted that when $\phi_{x_1,x_1}(f) = \phi_{x_2,x_2}(f)$, the SCoT filter becomes a ROTH filter.

Eventually, the maximum likelihood (ML) weighting functions derived from [Saa96] and [BS97] are oriented to increase the accuracy of the estimated ITD by means of attenuating the signals fed into the cross-correlation function in spectral regions where the SNR is very low. Such weighting function is defined as

$$\psi_{ML}(f) = \frac{\phi_{x_1,x_2}(f)}{|\phi_{n_1,n_1}(f)\phi_{x_1,x_1}(f)| + |\phi_{n_2,n_2}(f)\phi_{x_2,x_2}(f)|} \tag{2.25}$$

in which $\phi_{n_i,n_i}(f)$ corresponds to the power spectral density of the noise $N_i(f)$. This method is known to be robust to noise, but its performance in real-world applications is relatively poor, because reverberation is not modeled in its derivation. The approach proposed in [RF04] integrates the reverberation in the noise calculation so that the noise spectrum is defined as

$$\phi^{rev}_{n_i,n_i}(f) = \gamma\phi_{x_i,x_i}(f) + (1-\gamma)\phi_{n_1,n_1}(f) \tag{2.26}$$

where γ is a gain to tune depending on the reverberation level.

All these approaches can also be applied to gammatone filtered signals, that better approximate the filtering action of the cochlea. Hence ITDs and the ILDs are computed from each frequency decomposition given by the gammatone filters.

Eventually one can notice the wide variety of approaches centered on ITDs, whether it be for estimating its value or for establishing a relationship to source azimuth. Actually ILDs are quite straightforward to estimate but on the other hand, are challenging to exploit. There is no trivial modelling directly linking the value of an ILD to an azimuth direction. Hence, the main path of improvement for sound source localization approaches consists in enhancing ITD estimations from which it is easier to infer the source azimuth. In the same vein, distance cues and elevation cues until now are rarely exploited since they cannot be related by a simple model to a spatial location. Most approaches using elevation cues are based on the spectral notches, and thus the learning of the HRTFs based on a specific pinna [KSK+05, RIJG08, HLSVL06]. As for distance estimation, a review proposed in [Rod10] based on the signal amplitude, spectral amplitude and the binaural cues, showed that coarse approximation of the distance (< 1 m) could be obtained. This study was, however, performed under controlled and static acoustic conditions. Some other approaches obtained coarse estimation of the distance by using the DRR. In [LC10], the authors proposed a distance estimation from a DRR obtained through an equalization-cancellation method [Dur63], while the author of [Ves09] proposed a learning method based on the magnitude-squared coherence of binaural signals.

2.3 Localization paradigms

The last segment of this chapter focuses on the problem of sound localization in the robotic context. A typical approach for artificial auditory systems consists in replicating mammals auditory system such as the one described previously. In robot audition, the wide variety of robotic platforms offers, fortunately, more perspective for processing auditory cues. As a result, there exists two principal paradigms for sound source localization. The first paradigm corresponds to binaural approaches that are to a great extent inspired by mammals auditory system. In this context, most of the approaches are generally bio-inspired and consider anthropomorphic robots endowed with two microphones. Hence these methods take advantage of the knowledge of sound source processing mechanisms described in Section 2.1 and 2.2. On the other hand, the array-based processing generally consider free-field array of microphones, for which a direct relationship can be established between the auditory cues and the location of the sound source. This approach draws inspiration from the signal processing and sensor fusion fields. Yet, these two different paradigms share a common theoretical root based on interaural cues, more specifically on ITDs.

2.3.1 Binaural sound source localization

2.3.1.1 Typical approaches

For practical reasons (embeddability, space, resources consumption...), as well as for biomimetic sensing that gives more naturalness to interactions, binaural localization remains a major topic of interest in robot audition. In this context, localization solutions are based on the short-time cross-correlation of the signals between two microphones, derived from the Jeffress architecture. The interaural cues, more specifically ITDs, are at the core of the binaural techniques. Nevertheless, the task is highly challenging due to the low number of sensors. Moreover, in realistic environments, the interaural cues are particularly influenced by the morphology of the robot through HRTFs and by the acoustic properties of the robot location (*e.g.*, reverberation, noise) as mentioned in Section 2.1.2.2 and in Chapter 1.2. Regarding the complexity of acoustical environments, most of the early approaches tackled the localization problem in static scenes through simulations or fully controlled environments such as anechoic chambers.

Typically, these studies are based on the knowledge of HRTFs, to determine the azimuth and the elevation of the sound source, in controlled conditions. In [HLSVL06] or [KND06], localization is performed through the analysis of given HRTFs in order to link the azimuth and elevation to auditory cues. These latter works are principally based on ITDs and the spectral notches of the pinna. But these approaches are limited since they are fully based on HRTF knowledge: any change in the acoustic conditions, or in the binaural setup may severely degrade the localization performance. The simplified head HRTF models provides a flexible alternative, so that the localization does not depend on accurate measurements of the HRTFs. For instance, in [HAGK03], experiments of planar localization are successfully performed using the ST models in environments with moderate reverberation.

The RAEG has also been used in [NHOK02] or [NOK02] where complete solutions of localization, source separation and recognition are implemented. Yet, as expected the accuracy of such methods are decreased compared to HRTF-based localization techniques.

With the growing interest for robot audition, a wide variety of approaches have been proposed in order to cope with realistic conditions. The major problems tackled in the literature are: the flexibility of the localization performance, the robustness to acoustic conditions, the motion of the robot and/or the sound source(s), the presence of multiple sources in the scene. Such adverse conditions motivate the robot audition community to consider learning-based approaches instead of modelling all these conditions. In this context, because of the frequency-dependent pattern of ITDs and ILDs that vary considerably across robot platforms, azimuth-computation models are generally based on a prior training or calibration with a binaural setup. For more flexibility, the authors of [RVE10] shows that the azimuth can also be estimated by jointly taking into account ITDs and ILDs in a parametric model relating the azimuth angle to the interaural cues from pre-measured HRTFs. Unsurprisingly, this approach is less accurate than HRTF-based methods, but is able to cope with different binaural setups (*i.e.*, unknown HRTFs). However, the latter approach has been validated only in ideal simulated conditions. In more realistic conditions, the authors of [MvdPK11] introduces a probabilistic model (Gaussian mixture model) characterizing the complex interaction of ITDs and ILDs by training a model under various acoustic conditions to gain robustness. This method can cope with multiple sources and reverberation, but has also been validated in simulate environments. Likewise, in [WW12], from the knowledge of HRTFs, the authors propose a supervised learning integrating noise and reverberation to HRTFs to cope with adverse conditions. For a more practical purpose, in [DHSG15], the auditory cues are directly mapped to the spatial location of the source through a learning process in real world conditions. This learning process characterizes the related transfer function (RTF), the ratio of the TF from the left and right microphones, that is mixture of HRTFs and acoustic perturbations (see (2.7)). Hence, the method is limited to static configurations and to dedicated position in the environment: any change in the position of the robot or in the experimental conditions may degrade the localization process. For more flexibility towards changing acoustic conditions, [YAZ13] proposes a learning approach based on neural a network. However this method requires a training stage in several and various conditions to obtain accurate auditory cues in changing acoustic conditions.

Globally, modelling all variability of acoustic environment and/or perception, that is unique for each robotic setup, remains a complex problem to solve. Hence, most of the aforementioned methods generally consider static scenes, where the source and the robot are still in the scene at a favourable distance. Integrating the motion of the robot or the source in the localization process increases the degree of complexity of such approach as already outlined in Chapter 1.2.

2.3.1.2 Active audition

With the recent progress in the understanding of the human auditory system, a renewed interest in binaural techniques has raised in the robot audition community through *active audition* [NOK00b], that exploits the motion of the robot. The idea behind this approach resides in exploiting the motion of the robot to gather additional information for sound localization. The "naive" approach of *active audition*, simply consists in a triangulation of the measurements from several poses, as performed in [KB08, HST+99]. Obviously, in this context, the source must be static in order to obtain consistent results. Plethora of studies [LC11, MMWB15, TA06] show that the localization is performed more accurately with a motion than in a static way. In [LC11], the authors compared several motion strategies for the localization and concluded that the more beneficial strategy was obtained with purposeful motions, that consists in moving towards the sound source in order to reduce the distance. In [MMWB15], the comparison of different head motions shows that the best performing head movement is to rotate the head towards the most likely source direction. These results were quite expected, since in the first case, reducing the distance to the sound source increases the DRR level. Consequently, the interaural cues accuracy is increased, since the contribution of reflections is reduced in the recorded signal. In the second case, turning the head towards the probable sound source increase the spatial resolution of the localization compared to eccentric positions.

In realistic environments, spurious measurements are likely to happen and should be resolved. To cope with measurements uncertainties, the framework proposed in [KNM15], formulates the *active audition* as a tracking of bearing measurements problem. In this work, the data association problem is solved by associating a particle filter to an interaural coherency parameter to detect the direct sound in the recorded signal spectrum. Subsequently, the source location is inferred from multiple poses of the robot. In [PDA12], the authors propose to solve these issues by combining a stochastic probabilistic data association, for false measurements discarding, to an unscented Kalman filter. A potential intermittent moving source can then be tracked during a motion of the robot. This method has been applied so far in a simulated anechoic environment. A partial experimental validation is proposed in [PBD+14]. Recently, [NCVC16] extended this approach by integrating an estimation of the source activity from bearing measurements and a voice activity detection algorithm. This solution provides promising results able to cope with moderate reverberation, however the validation is still performed in simulation.

The aforementioned approaches consider a random motion of the robot that might not be favourable to sound localization. *Active audition* also covers purposeful motion strategies aiming to position the robot with respect to the source location. In this context, the motion of the robot is set in order to take advantages of the redundancy of the auditory cues measurements. In general, the motion strategies adopted by this kind of approaches are determined beforehand by the operator through heuristic approaches or fixed patrolling path. This is illustrated by [MEW09] where an incremental control strategy coupled with a neural network is used to position a robot closer to one static sound source. Similarly [TTLJ15] proposes a heuristic approach

combining ITD cues and range cues to home towards a sound source. However this kind of approaches face some limitations. First defining a strategy of motion beforehand can lead to unnatural behaviour of the robot for an interaction task. Furthermore, the heuristic approach requires to model and predict all the possible sensory state of the robot. In realistic environments that are essentially unpredictable, such approach would naturally be limited, while requiring the design of complex state machines. In a more sophisticated framework, the authors of [BPDCG12] develop a supervised learning based on sensorimotor theories, where an agent learns the relationship between the auditory cues and the state of its motors and move accordingly to the recorded map. The experimental studies associated to this work were performed in anechoic conditions [BNP+10]. Purposeful motion strategies can also be used to improve the localization performance. [VSC15] proposes a strategy to find the minimal motion that could lead to the localization through occupancy grids, while recently, [BDFP16] developed an information-based feedback-loop in order to minimize the uncertainty of the localization. Prior to this work, feedback-loop in robot audition context were introduced in [NMS02] and [KSM+03, KSK+05]. In the first work, a robot learns the acoustic map space (*i.e.*, relation between ILDs, ITDs and azimuth and elevation) and uses a feedback-error learning schema to orient the gaze of the robot towards the sound source. In the second set of work, the authors present a control scheme derived from a cost function characterizing the relationship between the position of a source and the orientation of a robot head. Although the cost function is empirically computed, the robot was able to turn towards the sound source without HRTFs measurements, in anechoic conditions.

Active audition compared to static localization, widens the perspective for sound source localization. From a personal point of view, *active audition* is a necessary step towards dynamic, thus natural, sensing in fluctuating environments. This approach reflects better the auditory sensing of living being, that is a dynamic process as expounded in Section 2.1.2.3. However when facing realistic environments, most of *active audition* approaches still lack of robustness as underlined by the numerous works performed in simulation or controlled environments. Indeed, the auditory cues are measured in a static way and still requires a good modelling of the acoustic conditions and the robot structure. Furthermore, such approach raises new challenges such as the fusion of uncertain measurements, the simultaneous motion of the sound source and the robot, and the motion strategy to adopt with respect to robot purpose. In view of these limitations inherent to binaural localization, array-based localization can be seen as an alternative.

2.3.2 Array-based sound source localization

Since the early nineties, a number of standard and highly functional methods based on microphone arrays have matured. Today, these sound source localization methods are used throughout numerous fields. Indeed these techniques are not necessarily limited to the field of robot audition but are also used for field such as tomography, telecommunication, seismology, or underwater acoustics. These techniques are mainly based on the developments in signal processing and sensor fusion. The speci-

ficity of robot audition compared to the latter fields comes from the the constraints described in Chapter 1.2 (real-time processing, embeddability ...). Nonetheless, compared to binaural processing, array-based approaches generally provide more accurate and robust results. The key advantage lies in the redundancy and spatial diversity of the signal measurements conversely to static binaural methods. Furthermore, depending on the microphones topography, the global distance, elevation and azimuth angles can also be inferred by simple triangulation of the different local azimuths computed from each pair of microphones. The array-based techniques principally focus on the computation of the delay of sound arrival between the different channel pairs. This feature also known as the time difference of arrival (TDOA), TDE (time delay estimation) or interchannel time difference, is fairly equivalent to ITD in AEG context. Hence these methods rarely consider head-mounted systems, but rather array of free-field microphones under the far-field assumption. We are attached to described the main localization techniques in array-based sound processing in robotics. The techniques described below are classified in three general groups, based either on triangulation solutions, beamforming or MUltiple SIgnal Classification (MUSIC).

2.3.2.1 TDOA-based approach

From the binaural approach described above, the most straightforward approach consists in triangulating the TDOAs of several pairs of microphones. The TDOA-based approach is the transposition of the binaural localization through interaural cues, to arrays context. This method can be categorized as an indirect approach that usually follows a procedure in two phases. First the estimations of the TDOAs between microphone pairs is performed and, afterwards, the source position is estimated based on the geometry of the array and the computed TDOAs. For simple microphones topography, a basic geometric rule allows to infer the sound location from several source azimuth obtained from TDOA measurements. For more complex array configurations, the sound localization problem becomes an optimization problem, where the actual sound location minimizes a set of equation related the measurements. Under the far-field assumption, for a given pair of microphone $\mathbf{M_1}$ and $\mathbf{M_2}$, the TDOA τ with respect to a source $\mathbf{X}_s$ is defined as

$$\tau_{1,2} = \frac{\|\mathbf{X}_s - \mathbf{M_1}\| - \|\mathbf{X}_s - \mathbf{M_2}\|}{c} \tag{2.27}$$

arbitrary considering $\mathbf{M_1}$ as the reference. As proposed in [VMRL03], when considering N microphones, a system of $N-1$ equations has to be solved, leading to a system where

$$\mathbf{M}\mathbf{X}_s = \mathbf{D}. \tag{2.28}$$

In the latter equation $\mathbf{M}$ is a matrix referring to the microphones position with respect to a reference, that is composed of the set of vectors $[\mathbf{M_i} - \mathbf{M}_{ref}]$. As for $\mathbf{X}_s(x_s, y_s, z_s)$, with $z_s = 0$ for planar scenes, it denotes the sound source position. At last the vector $\mathbf{D}$ is defined as $\mathbf{D} = [c\tau_{ref,i}]$. In a planar scene the system requires at least 3 microphones, while for 3D localization at least 4 microphones are required

to solve (2.28) through

$$\mathbf{X}_s = \mathbf{M}^{-1}\mathbf{D}. \tag{2.29}$$

When more microphones are available, the system is overdetermined and the solution can be found using the pseudo-inverse

$$\mathbf{X}_s = \mathbf{M}^{+}\mathbf{D}, \tag{2.30}$$

which can be computed only once since $\mathbf{M}$ is constant. However the spatial resolution of this approach is generally poor, and the localization is degraded by reverberation when several spurious TDOA $\tau_{ref,i}$ are extracted. TDOA-based approaches are also dependant on the spatial configuration of the microphones in order to avoid singularities of the matrix (*e.g.*, planar array configuration for 3D localization). And eventually this type of approach is not adapted to multi-source localization, for which data association problems should be solved.

TDOA-based approaches due to their simplicity dominate most of the early contributions in robot audition. The sound location is simply deducted by triangulation of the TDOAs attached to each microphone pairs. This idea was successfully exploited in [HST+99], for building an navigation system based on hearing perception on a mobile robot. In [VMRL03] the authors presented a 3D localization system from the triangulation of the TDOA obtained by a cubic-shaped array of eight microphones. Although MUSIC and especially beamformer are nowadays commonly used, TDOA-based methods are still applied for localization principally because they are intuitive, and allow to fully localize a source (azimuth, elevation and distance) with the appropriate microphones topology. In this context [KCCL08] uses three microphones to extract the sound location from a GCC. In a recent work [APH14] uses a branch and bound optimization for 3D localization from an arbitrary non-coplanar microphones array topography. In [MP10], the authors propose to infer the sound location by combining each extracted TDOA into a mixture of von Mises distributions. The PHAT processor that is known to be particularly robust to reverberation is also used in many contributions such as [LSWL12] in which the authors integrates a denoising function into PHAT processor, for self-localization as depicted in [HCW+11] with an acoustic SLAM framework or in [BIK+15] where TDOAs are used to estimate the 3D pose of a hose-shaped robot.

However, these approaches are generally limited to localize a single source. Indeed, the number of active sound source should be known beforehand since it cannot be reliably determined from the cross-correlation function. The peaks in the correlation function could also be caused by reflections of the sound. And more importantly the problem of associating a given TDOA to one of the active source should be solved in order to obtain consistent localization results.

2.3.2.2 Beamforming-based approach

The second approach widely used in robot audition is beamforming [VVB88]. Beamforming techniques consist in combining the signals measured by an array of sensors, in order to focus the latter signals on a specific location or direction. In machine

hearing context, the signals recorded by a microphone array can be steered over a direction that contains the actual sound source. Typically, the signals recorded by each microphone are filtered separately so that the absolute energy of the output peaks out at the source location. Since microphones in the array are spatially separated, the source signal will arrive at each microphone at a different time. Hence considering N microphones recording and S sound source the output of the beamformer is given as follows:

$$B(f,r) = \sum_{s=1}^{S} \sum_{i=1}^{N} W_i(r,f) X_s(f) e^{-2j\pi f \tau_{i,s}} \tag{2.31}$$

in which W_i express the filtering action, $\tau_{i,s}$ the delay of arrival of the signal s on the microphone i from an arbitrary reference point, and r the spatial direction of research. In practice, (2.31) is evaluated on a limited set of potential directions, so that generally r corresponds to a sphere for 3D localization or a circle for planar localization. For instance in [VMR07], the authors defined the directions of research on a sphere build upon icosahedral grid that sweeps all the research space for 3D localization. Eventually, the output power is obtained by integrating (2.31) over a time window.

Inside beamforming techniques, a set of methods are based upon analysis of spatio-spectral correlation matrix derived from the signals received by the microphones. In this set, one can cite the Minimum Variance Distortionless Response beamformer (MVDR), that is widely used in signal processing and perform well in noisy environments. However, in robotics more conventional beamformer are used for real-time purpose. One of the simplest form of beamforming used in this context is certainly the *delay-and-sum beamformer* (DSBF), [JD92], in which the filters follows a delaying pattern. In this case, for a given sound source, the output of the beamformer is maximal when the filters delays τ_i between each pair of microphones so that the microphone signals are in phase, and therefore add constructively. Hence the DSBF adds appropriate time shifts to the microphone signals to compensate for the propagation delays. Thus, a given filter $W_i(r,f) = e^{2j\pi f \tau_r}$ delays the signal x_i by τ_r that fits the direction r. In machine hearing context, the beamformer relies on the results of the cross correlation between each pair of microphones (see (2.18)) that are delayed and summed so that for a direction r

$$B(r) = \sum_{i,j=1}^{M} \boldsymbol{R}_{i,j}(\tau_{i,j}(r)) \tag{2.32}$$

in which i,j denotes the microphone pair and M the number of microphones. By integrating (2.32) over several a set of research direction, the sound localization can be inferred from the peaks of the absolute energy of (2.32). This approach is simple and requires a low computation cost, which explains that this is the most exploited technique for array-based localization. When the PHAT weighting function (see (2.22)) is used in (2.32), the beamforming technique is generally assimilated to the well-known SRP-PHAT algorithm.

The convincing results deployed in [VMHR04, VMR07] certainly popularized the DSBF. In this work, implemented in *ManyEars* software, the azimuth and elevation

of multiple moving sources are estimated from eight microphones, in real-world environment. The low computation cost of the approach also favoured the utilization of such method. As a direct consequence, approaches based on huge amount of microphones arises for robust localization in realistic environment. These approaches are illustrated by [SKM06] that is based on 32 microphones in order to self-localize the robot with respect to auditory landmarks, or by [SKT+12] in which a spherical array of 64 microphones is used to localized sound sources in dynamic environments. In [NKD+09], the authors considered an orientation-extended amplitude beamforming in order to infer sound location and direction from an array of 96 microphones. Globally all the aforementioned approaches suffer from poor accuracy at low frequencies and assume far-field conditions. In the near-field the localization performance is drastically reduced. Thence [ADS06] proposed to extended the beamforming to near-field conditions under the knowledge of the distance of observation of the source.

Beamforming techniques suffer from the limitations inherent to ITDs, that is to say limitation at high frequencies because of the aliasing of ITDs for short wavelengths (see Section 2.1.2.1). On the other hand, at low frequencies, the beamformer loses accuracy, the energy peaks are very wide as exposed in [DZD01], which means that the spatial resolution is poor. Moreover beamforming methods can only infer azimuth and elevation angles from signal measurements, unlike TDOA-based methods the distance to the source remains unknown. Thence beamforming techniques would be particularly suitable for wide arrays that allow to cope with high frequencies without aliasing.

2.3.2.3 MUSIC-based approach

At last MUSIC [Sch86] is also commonly used in robot audition to infer sound localization from an array of microphones. In its original form, MUSIC goal is to estimate from measurements a set of constant parameters characterizing the received signals at a frequency f. In robot audition context, the parameters to be estimated are the number of emitting sound sources and their position. In plain, the method perform an eigenvalue decomposition of the correlation matrix of the recorded data, in order to split the noise subspace to the signal subspace from which the source(s) parameters can be inferred. Assuming a data matrix $\mathbf{X}(f) \in \mathbb{C}^{M\times N}$, with M the number of snapshots of the array signal and N the number of sensors, $\mathbf{X}(f)$ is expressed as follows:

$$\mathbf{X}(f,t) = \mathbf{A}(f)\mathbf{S}(f,t) + \mathbf{W}(f,t) \tag{2.33}$$

where $\mathbf{A}(f) \in \mathbb{C}^{M\times D}$ is matrix built upon the steering vectors characterizing the delay of arrival of a given signal, while $\mathbf{S}(f) \in \mathbb{C}^{D\times N}$ is the incident source matrix characterizing the sound source(s) received by each microphones, and $\mathbf{W}(f) \in \mathbb{C}^{M\times N}$ the noise spectrum. For simplicity, we omit in the following development the frequency and time dependencies. Generally $A(f)$ is determined beforehand and characterizes the research space of the sound location. Under these conditions, the correlation matrix of the $\mathbf{X}$ is then

$$\boldsymbol{R}_{x,x} = \mathbf{E}[\mathbf{X}\mathbf{X}^H] \tag{2.34}$$

where the function $\mathbf{E}[\,]$ expresses the expectation, and $.^H$ characterizes the Hermitian transpose. From (2.33) the latter expression can be developed as

$$\boldsymbol{R}_{x,x} = \mathbf{A}\mathbf{E}[\mathbf{S}\mathbf{S}^H]\mathbf{A}^H + \mathbf{E}[\mathbf{W}\mathbf{W}^H] = \mathbf{A}\boldsymbol{R}_{s,s}\mathbf{A}^H + \boldsymbol{R}_{w,w} \tag{2.35}$$

If the noise is not correlated with the sound signals, (2.34) can be simplified as

$$\boldsymbol{R}_{x,x} = \mathbf{A}\boldsymbol{R}_{s,s}\mathbf{A}^H + \sigma_w^2\mathbb{I} \tag{2.36}$$

The eigenvalues of the covariance matrix are then obtain by solving

$$|\boldsymbol{R}_{x,x} - \lambda_i\mathbb{I}| = 0, \tag{2.37}$$

which can be rewritten using (2.36)

$$|\mathbf{A}\boldsymbol{R}_{s,s}\mathbf{A}^H - (\lambda_i - \sigma_w^2)\mathbb{I}| = 0, \tag{2.38}$$

Since $\mathbf{A}$ contains steering vector linearly independent and $\boldsymbol{R}_{s,s}$ is not singular for non correlated sound signals, when there are less source than microphones (D<N), $\mathbf{A}\boldsymbol{R}_{s,s}\mathbf{A}^H$ is positive semi-definite of rank D. This result automatically implies that M-D eigenvalues are null, then $\boldsymbol{R}_{x,x}$ has M-D eigenvalues corresponding to noise variance σ_w^2. By sorting the eigenvalues of $\boldsymbol{R}_{x,x}$, the largest eigenvalues span the signal subspace while the lowest span the noise subspace. The number of sound sources is then determined by $M - K$, K referring to the number of smallest eigenvalues corresponding to noise. The steering vectors corresponding to the source locations lie in the signal subspace and are hence orthogonal to the noise subspace. By searching through all possible array steering vectors to find those which are perpendicular to the space spanned by the non-principle eigenvectors, the source location can be estimated. Then the DOAs of the multiple incident signals can be estimated by locating the peaks of a MUSIC spatial spectrum given by:

$$P(\alpha) = \frac{1}{\mathbf{A}^H(\alpha)\boldsymbol{\mathcal{N}}\boldsymbol{\mathcal{N}}^H\mathbf{A}(\alpha)} \tag{2.39}$$

where $\boldsymbol{\mathcal{N}}$ is a matrix containing the eigenvectors obtained from the noise subspace. Actually, the orthogonality between the noise subspace and the signal subspace characterized by the steering matrix minimizes the denominator of (2.39) and hence gives rise to peaks in the spectrum. Nevertheless, in practice $\mathbf{X}$ is not fully known, since only one recording is available (*i.e.* $M = 1$). In general, $\mathbf{X}$ is approximated over a time length L so that

$$\widehat{\boldsymbol{R}}_{x,x} = \frac{1}{L}\sum_{l=0}^{L-1}\widehat{\mathbf{X}}_l\widehat{\mathbf{X}}_l^H, \tag{2.40}$$

where $\widehat{\mathbf{X}}_l$ approximates $\mathbf{X}$ from a M-point Discrete Fourier Transform on the time snapshot l. Subsequently an estimation $\widehat{\boldsymbol{\mathcal{N}}}$ can be extracted from (2.40) in order to compute (2.39). This method gives generally accurate estimation of azimuth and/or elevation angles but in exchange of a heavy processing payload. Moreover MUSIC strongly depends on the estimation $\widehat{\mathbf{X}}_l$ that conditions the performance of

the localization. Ultimately, such methods is designed for narrowband sound sources, since the result obtained is valid at a given frequency f. As a result several extensions are proposed in robot audition literature in order to fill these gaps.

MUSIC was introduced in robot audition in the early 2000s. In [AAM99], the authors proposed an extension of MUSIC to cope with broadband signal. The approach consists in splitting the signal in narrowband frequency and computing the average spectrum over all frequencies. To cope with noise it is proposed in [NNA+09] to integrate a noise correlation matrix in the estimation of the data matrix (2.40). This solution improves the differentiation between the signal eigenvalues and noise eigenvalues in presence of loud noise. The noise correlation matrix being freely tunable, the authors of [OYNN12] developed an incremental estimation of the noise matrix by averaging the current noise matrix with the past estimations of the matrix over a given time window. In the context of sound localization from aerial vehicles that produce loud and dynamic noise caused by propellers, [ONM+14] improves the latter method by introducing a correlation matrix scaling to adapt the noise reduction level dynamically, while in [FON+13] the noise correlation matrix is computed from the motion information obtained by the Inertial Measurement Unit (IMU) of the quadrotor. However the main flaw of these methods still remains the computational cost of MUSIC methods. Most of the aforementioned approaches are either performed only on a single frame or validated offline, since they do not meet real-time requirement. Hence, in order to overcome this limitation, in [NNI12] and [NGN13], for 3D localization, proposed to use a singular value decomposition (SVD) instead of the eigenvalue decomposition that is more demanding computationally, coupled with a hierarchical search strategy. In the first contribution the approach performed 3 times faster than baseline MUSIC methods, while for the 3D case the computational cost was reduced by a factor 30.

2.3.2.4 Sound source tracking

Independently of the approach used, the gain of robustness of the array-based approach allowed for several methods addressing the problem of moving sound sources. Being able to accurately follow a moving sound source is a significant challenge in robot audition community. When considering an unknown motion of the sound source(s), erroneous localization are more complex to detect and correct, in realistic environment. Additionally the motion limits the window length from which the localization is performed, which has proven to be important for accuracy (see Chapter 1.2). A straightforward approach consists in continuous static localization. This approach obviously requires real-time performances in order to process the sound accurately. In [SJHS15], an heuristic method based on a particular shape of four microphones is proposed to track one sound source. To reduce the computational cost of the algorithm, the authors based their tracking on the sign of the extracted TDOAs.

More sophisticated tracking have also been envisioned by introducing filtering methods in the localization process to keep the track of the sound source(s). However only few works address the tracking of dynamic sources, in real world conditions.

In [WL03], the authors introduce a particle filter framework which could improve sound tracking when using cross-correlation methods or beamforming in moderately reverberant room. This type of approach is also exploited in [VMR07] where a solution including eight microphones and a particle filter is able to track multiple sound sources from a DSBF. Similarly the method proposed in [MP10] tracks one sound source by using four planar microphones with a particle filter while in [EMN+15], a probability hypothesis density tracker is proposed. In [KCCL08] with a short signal window length and a Kalman filter, a robot is able to track a moving sound source by triangulating data from cross-correlation results.

Nonetheless, the relative robustness of the array-based methods implies several counterparts. In most of the aforementioned work, the robustness is gained from the data redundancy. The more microphones in a wider array, the more robust the localization system. Hence, these techniques generally imply a higher computation cost with more data to treat, that can challenge the real-time requirements. In the same way, the embeddability on robots, especially on humanoid robots, limits the number of microphones and the available space, which can be incompatible with the array-based methods.

These are the main localization techniques used in robot audition when the sound is acquired from an array of microphones. However, we just reviewed the basic theory of array processing for localization. Other classifications and techniques might be found in signal processing literature. Usually, in signal processing, angular spectrum methods and spectral clustering methods define the categories of localization techniques. Spectral clustering methods consist in iteratively estimating the time-frequency bins associated to each source and update the source TDOAs according to the observed phase differences, by clustering algorithms. In the angular spectrum approach, the objective is to build a function of TDOAs that is maximized at the actual source azimuth. This kind of approach includes beamforming techniques. A typical angular spectrum technique, in robot audition, is the maximum likelihood estimation proposed [PBDM14] in the context of binaural localization though. Readers interested in extensive studies are referred to [BOV12, BVMA09] which provide comparison of different methods under varying conditions (reverberation, active sources, noise...) and [ADS15] that provides a comprehensive review of these techniques in robot audition. For a thorough study addressing array processing in a general context, a textbook such as [HT02] is a reference.

2.4 Conclusion

Endowing robots with a reliable hearing sense, and more particularly a sound localization framework, is a key feature for functional auditory scene analysis. The challenges raised by the robotic context, such as real time constraints, dynamic localization (motion of the source and/or the robot), lead the robot audition community to take inspiration from human auditory system. Still, the human auditory system remains complex, and all the mechanisms involved in the auditory sensations are not

fully understood. Among the auditory cues, the interaural cues, more particularly the interaural time difference, have been identified at a very early stage as related to the spatial location of the sound, and focused most of the studies in psychoacoustics. However, it is clear that the localization process involves the interaction of several cues in complex ways, that for now cannot be fully explained.

In addition to interaural cues, frequency-dependent filtering by the body also known as HRTFs play a role for localization in the vertical plane and for resolving *front-back ambiguities*. The distance is suspected to be inferred from the ratio of direct sound over reflected paths and from the sound level. Unfortunately, until now, no simple auditory model can relate these cues to a spatial location of the source even though their contributions to localization have been acknowledged.

The prominent contributions of sound localization in robotics are based on the interaural cues in a binaural context. The use of such cues is, however, limited by the variability of the sound perception, characterized by the acoustic conditions and HRTFs in realistic environments. The binaural localization requires an accurate modelling of these parameters, which is a significant challenge. As a result, most of the models and techniques proposed in the literature consider static scenes, and are either validated in simulation or experimentally in anechoic conditions. Learning methods have also been explored in order to cope with realistic environments. But these methods generally require a large training sets and complex learning procedures. Furthermore the estimated auditory cues generally correspond to a particular environment, with a given microphones configuration or robot structure. These methods are hardly adaptable to unknown environments since estimating all possible combinations of acoustic configuration is not yet affordable. Thus, solutions considering arrays of microphones have also been explored in order to improve the robustness of the localization process. Although this approach generally gives better results, the computational cost, the size of the array, can be crippling in embedded artificial auditory system, while the performance for dynamic scenes still remains mixed.

More recently, *active audition* has been designed to overcome such limitation by using the motion in order to benefit from measurements from different robot poses. The motion strategy to develop remains a central point of the *active audition*. From different studies it appears that purposeful motions (getting closer or turning in the direction of the source) increase localization performance. These results are also corroborated by recent psychoacoustics studies that characterize the perception as a dynamic process where self-initiated motions are connected to the varying acoustic cues. However the approaches developed so far based on heuristic algorithms or fixed patrolling techniques, do not provide enough flexibility. Indeed *active localization* still requires a modelling of the perturbations, that may be challenging to obtain in dynamic and realistic environments.

A different path, proposed in this thesis, consists in considering a low-level connection between the motion and the auditory perception, in order to avoid explicit localization and modelling the perturbations. The work depicted here considers a generic approach that couples the motion of the robot to the auditory perception in a closed-loop control. The control loop moves the robot directly from the auditory

feedback, instead of focusing on the preliminary explicit source localization that is the common path of improvement taken by the methods reviewed above. In these methods, the task generally consists in locating the sound source first and move the robot afterwards in an adequate position: the auditory system is completely independent from the control of the robot. By contrast, our approach is centered on the connection between auditory perception and robot control. This connection can be expressed through a sensor-based control framework detailed in the next chapter.

Chapter 3

Basics of sensor-based control framework

This chapter divided into two main parts introduces the principles of sensor-based control that build up the theoretical aspects of this thesis. The purpose is to provide an alternative to sound source localization in realistic environments for robot positioning tasks.

In the first part, we introduce our formulation of robot motion control based on acoustic cues, that is an approach centered on robot control rather than sound localization. Thence, before a thorough analysis of the theoretical properties of sensor-based control, the benefits of using such a feedback loop for robot audition are exposed in Section 3.1.1 through a typical positioning task based on auditory information. In Section 3.1.2, a short review is given about the main usage and applications of sensor-based control in robotics.

The second segment of this chapter focuses on the theoretical aspects of sensor-based control. The concept of task function leading to the principle of sensor-based control is first introduced in Section 3.2.1. Following this study, the actual theoretical framework that builds up *aural servo* (AS) paradigm is presented in Section 3.2.2, while accounting for the stability of the controller in 3.2.3 and for the task to be performed by the robot, as exposed in Section 3.2.4.

3.1 Introduction to sensor-based control

3.1.1 On the interest of sensor-based control for robot audition

What would bring sensor-based control to robot audition topic?
In this section we try to give a synthetic answer to the latter question by emphasizing the properties induced by such paradigm. Such properties greatly motivated the realization of this thesis. The previous chapter pointed out the limitations of sound source localization that prevents robot audition applications from evolving in realistic environments. Until now, especially in sound source localization, most of the contributions tackled the issues of robot auditory perception from acoustic, signal processing or physiological perspectives. Typically robot audition is addressed

in the same way as machine hearing. It is not uncommon to apply robot audition techniques to other types of machine. This is the case of the framework reported in [NSN15] that uses *active audition* to localize in 3D a sound source from a tablet device. Likewise, the human-machine interface proposed in [NMN15] uses the developments of robot audition for a car navigation system. In this paradigm commonly adopted by the robot audition community, robots are considered as application platforms, since most of the solutions developed are not exclusive to robots. Only the context of robot audition differs from machine audition, since the latter solutions should account for additional constraints such as ego-noises, real-time processing, limited resources, already mentioned in Chapter 1.2.

Nonetheless, unlike other machines, robots are endowed with the motion ability. Robots can move in accordance with a given purpose, such as interacting with the environment. This simple characteristic can make a tremendous difference in auditory perception. Yet, the previous chapter stressed the benefits of motion for auditory perception through *active audition* framework that could increase the robustness of sound source localization. However the main flaw of the latter framework comes from the predefined behaviour of the robot that cannot account for the perturbation and the changes in the environment. *Active audition* as well as classic localization techniques primarily focus on the quality of auditory perception from an acoustic and signal processing perspective. These techniques are attached to solve the question *Where is the sound source ?* that corresponds to a machine hearing problem in the broad sense. The control of the robot and the concept of giving a purpose to the robot are generally neglected. As a result coping with realistic (*i.e.*, dynamic) environments becomes challenging since no feedback from the environment is used.

By contrast in this thesis, we aim to develop a dedicated control framework for auditory perception, that takes advantage of the characteristics of robots over machines. Indeed control, as an essential component of robotics, is introduced in the robot audition topic. Thus AS paradigm exposed in this manuscript is particularly centered on robots, as a dedicated control framework for robot audition. The previous questioning about the source location, that builds up the sound source localization paradigm, becomes *How to position the robot with respect to the sound source?* The latter way to consider the auditory interaction can only be solved from a robotic perspective, since it requires a voluntary motion of the robot. In this paradigm, we thus define the auditory perception as a task to be completed by the robot. In this way, instead of only considering the constraints inherent to the robotic context, we also exploit the advantages of robot compared to other machines. More explicitly, our approach consists in defining the motion of the robot with respect to the auditory feedback measured from the microphones. Typically this is the principle of the sensor-based control approach. To emphasize the interest of expressing the auditory perception in this paradigm, let us detail a simple example of auditory task. This task is a basic head-turn strategy in a planar scene. More exactly, the task consists in orienting a robot head towards the given direction of the sound source. In this scenario, the robot is endowed with two microphones, while the sound wave emitted by a source reaches the microphones with an angle α as illustrated in Figure 3.1.

In a classic approach, the task consists in extracting the interaural cues (*e.g.*,

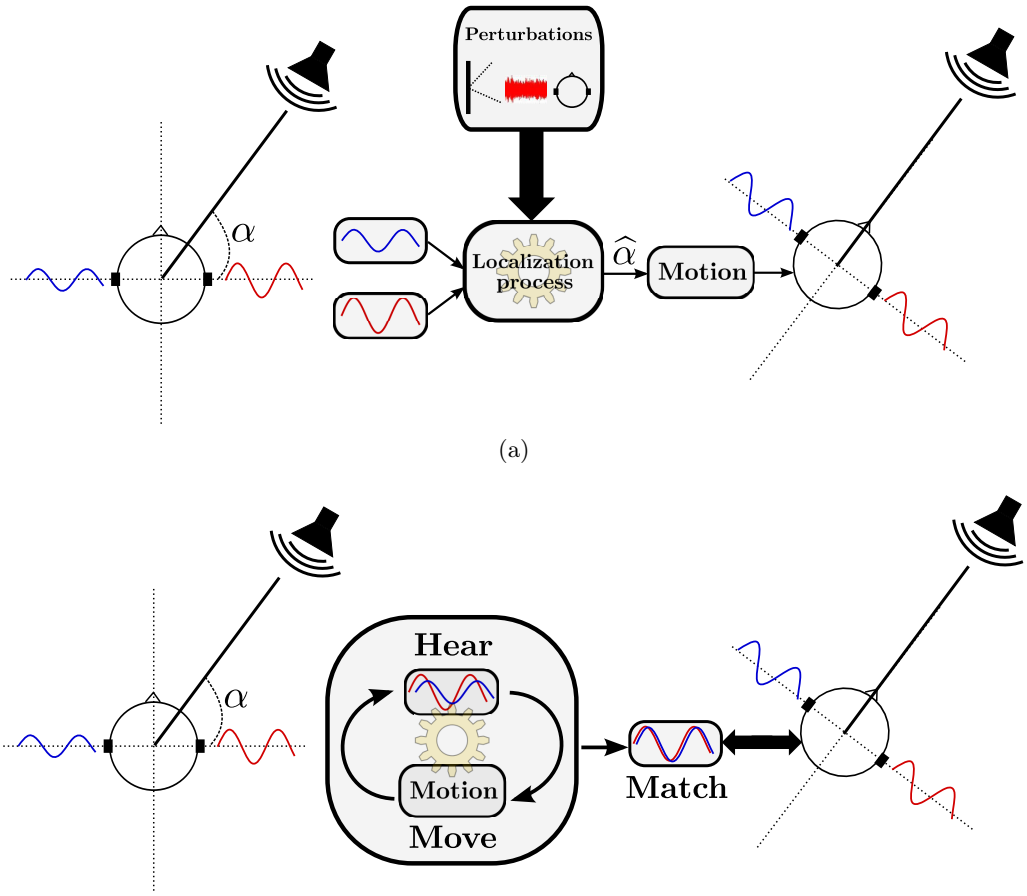


Figure 3.1: A robot perceives a sound source emitting from an angle α. A head-turn task can be performed from two approaches. (b) Classic localization: After estimating the sound direction $\widehat{\alpha}$, the head is turned accordingly; (c) Sensor-based approach: The head is turned until the two signals recorded by the left and right microphones match, which corresponds to the desired orientation of the robot.

ITD), computing the corresponding orientation $\widehat{\alpha}$ and eventually turning the robot head by $\widehat{\alpha}$. The good proceeding of the task depends on the accuracy of $\widehat{\alpha}$ regarding α, so that the robot faces the correct direction. This constraint implies an accurate modelling of the potential perturbations: reverberation, noises, shadowing effect of the head, ego-noises.... Taking into account all these parameters is challenging since they are for most of them uncontrollable, unpredictable and highly fluctuating. Furthermore if an accurate estimation of the perturbations is available, it is likely that the localization would be overfitting the acoustic environment.

In opposition to the latter approach, in a sensor-based control framework, the task consists in moving the head in the appropriate direction until the signal recorded from the left microphone matches the signal recorded from the right microphone. Having identical signals from both microphones logically implies that the robot is facing the

sound source. One immediate advantage of such task modelling lies in the fact that the sound source localization is not required anymore. Indeed the behaviour of the robot is not controlled by the estimated location of the source but rather steered by the auditory properties implied by the desired pose (*i.e.*, similar measurements from both sensors) and the dynamics of the sound features with respect to the motion. These two characteristics are essential key points that emphasize the robustness of the sensor-based approach over the classic sound localization.

First the dynamics of the sound features is an element that may be considered robust with respect to usual perturbations faced in robot audition. Actually the dynamics of the sound features with respect to the motion remains consistent, independently of the level of reverberation, ambient noise, or the HRTF of the robot. In Figure 3.1b, without any thorough sound analysis while achieving the task, it is expected that the amplitude of the right signal should decrease while the left signal amplitude increases. When the robot faces the sound source, these two signals should have the same amplitude. A similar reasoning holds for the delay of perception of the signal between the two microphones. Analogously, the sound level increases when the sensors moves closer to the sound source while it decreases with an opposite motion. This simple evidence has not been exploited yet in robot audition and certainly builds up the robustness and flexibility of the AS paradigm. The good achievement of a task depends on the consistence of the dynamics of the sound features and not on the consistence of their intrinsic values as it is for sound localization approach.

The second characteristic exploits the current auditory features with respect to desired auditory features. These desired features build up a reference that should be reached. They are either experimentally measured or estimated. The system is thus defined with respect to local measurements rather than global measurements as it is usually assumed for sound localization. Intrinsically, the desired features could be erroneous with respect to a global ground truth obtained from a localization perspective. However, as long as these values can be reached, the control task can be correctly achieved. When considering more specifically the context of facing the sound source, signal measurements around this pose are known to be the most accurate: the spatial resolution and the accuracy of the auditory cues are at their finest. This property has already been discussed for machine hearing in [CLLH07] reporting that the performance of localization and tracking algorithms can be improved by turning the head to ensure that the tracked source remains in the region directly in front of the head where the localization cues are most sensitive. The same evidence holds for human localization performance as exposed in [CLH97] or [MM90]. Consequently, in this context, the control framework can reach with a higher accuracy the desired auditory features.

Hence the combination of these two characteristics constitutes the key strength of formulating the motion control in a sensor-based control approach. Our formulation of the problem simply relaxes the accuracy requirements, that is a critical point for sound localization approach. As a consequence the effect of perturbations like reverberation and potential noise that alter the auditory cues estimation, are then limited regarding the proceedings of the task. A beneficial effect of this result concerns the well-known subjectivity of the auditory perception characterized by HRTFs. This

subjectivity is highly constraining for localization systems embedded in robotic head (as discussed in Chapter 2) and generally requires an accurate modelling of the perturbations induced by the robot structure. In the task represented in Figure 3.1b, the symmetrical property of the robotic structure (valid for most of the robots) implies that the shadowing effect from left and right microphones should be the same when the robot is facing the sound source. Moreover knowing that the dynamics of the auditory cues are not affected by HRTFs, the scattering effect of the head is nullified and the latter task can be completed without any modelling of HRTFs, as it will be shown in Chapter 6.

Eventually one can highlight the advantages related to closed-loop control scheme. The closed-loop control is a highly reactive process that is especially adapted to dynamic environments, since a new control input is given to the robot at each processing frame. Hence this method should be able to cope with moving sound sources. Likewise, the measurements are updated and corrected in real-time, which limits the effect of spurious and erroneous estimations that are not consistent with the motions of the robot. As stated above, the motions of the robot induce a given dynamic of the auditory cues that can be anticipated. And, similarly to *active audition* processing, several measurements bring additional information that might increase the robustness of the perception.

From all these observations, it appears that AS framework can bring robustness and flexibility to auditory tasks, which is a principal requisite for robots evolving in realistic environments. The potential benefits reported at this stage will be supported all along the development of this manuscript by theoretical and experimental evidence in Chapters 4, 5 and 6.

3.1.2 Sensor-based control in robotics

One of the most well-known sensor-based application is the so-called *visual servoing* (VS) that makes use of computer vision data to control the robot dynamically. VS refers to the paradigm of closed loop control techniques of actuated systems with visual feedback. With the high preference of vision sense over other modalities, VS naturally becomes a tool for automatic control in unknown, complex and dynamics environments. The first approaches of sensor-based control using vision were introduced in the late 70s [Agi77, Agi79] followed by [WSN87, RCE90, FM89] in the 80s. These works pioneered the automatic robot control based on visual information. Indeed despite the limited performance of the computing architectures in the 80s, these works allowed to exploit visual information in real-time with reactive behaviours.

Nowadays, regardless of the diversity of vision-based robot control that benefited from improvements in computational power, VS is still considered as an efficient approach for vision-based robotic tasks such as positioning and tracking. Based on the nature of the visual information, the existing VS approaches can be classified into two main categories: *image-based visual servoing* (IBVS) and *position-based visual servoing* (PBVS) [WSN87, CH06]. In IBVS techniques, visual features (*e.g.*, points, lines, moments) in the image plane are used to compute the control law. In PBVS techniques, the visual information is used to extract the pose of the target with

respect to the robot and the control law is computed from the error between the current and the desired poses. In AS, this approach would consists in localizing the source at each iteration of the control loop. However, because of the localization limitations discussed in the two previous chapters, this approach is not interesting (yet) for AS framework.

In recent decades, VS techniques have been widely applied to different platforms such as classic robot manipulator [C+96], mobile robots [Cor03], or more recently to aerial vehicles [MH07]. VS has already been applied to diverse application contexts. In medical robotics the accuracy inherent to VS is used to develop medical assisting robots. This is illustrated by plethora of contributions such as [KGD+03] that proposes a robotic system that automatically positions surgical instruments during robotized laparoscopic surgical operations. In [YN01], an automated embryo DNA injection system controlled within a VS framework is reported while ultrasound images are used to control an ultrasound probe held by a medical robot in [MKC08]. VS has also been applied to virtual reality field by formulating the problem of camera pose estimation as a regulation task achieved in the image plane [MC02, CMPC06]. More recently the visual feedback has been used to control the deformation of compliant objects by robot manipulators [NALRL14] through the *visual deformation servoing* framework. This wide range of applications classify VS approach as a general paradigm, that is shaped by the different contexts and purposes of the aimed task. Unsurprisingly VS has been used with different types of visual sensors beside the classic perspective camera. The utilized sensors could be catadioptric cameras [BMH03], generalized cameras [CMS11], stereoscopic [HCM94] or RGB-D sensors [TM12] among others, that could also be exploited while being externally fixed [HDE98].

Nonetheless the sensor-based approach is not limited to visual feedback and has been applied to a wide range and variety of sensors. In the early nineties, [EMS90] showed that most sensor-based applications may be treated within a unified control loop framework. In the latter paper, the authors develop a global approach to the problem of proximity and force-based control in robotics applications within a sensor-based control framework that lead to the development of *proximity servoing*. Nowadays a significant number of applications using sensor-based control are related to touch sensing feedback, whether it be human-robot interaction for collaborative tasks, obstacle avoidance, manipulation or grasping tasks. This can be illustrated by [KJHJ05] in which the authors introduce a framework to control a team of robots through proximity sensors (*e.g.*, laser) feedback that follow walls or terrains at a given distance. The application proposed in [WS08] explores the use of electric field pre-touch as a feedback signal for closed loop control for aligning and pre-shaping a robotic hand for grasping tasks. The paper proposed by [SZS94] showed that sensor-based control could also be used with the tactile images produced by tactile sensors mounted on the fingertips of a robot hand for object manipulation. This paper gives rise to the well-known *tactile servoing*. Tactile feedback can be used nowadays for tactile exploration tasks [LSHR13] or for more complex tasks, like dexterous prehension as proposed by [OSC00].

Furthermore, the versatility of sensor-based control framework engendered the

topic of multimodal sensor-based robot control. In this context, a task is achieved by combining data of sensors measuring different physical phenomena in an unique control framework. Combining several sensors information in the sensor-based approach can be interesting to overcome the limitations and flaws inherent of a particular sense (*e.g* limited field of view of cameras) by using complementary modalities, or for achieving more complex tasks that can be related to human-robot interactions or object manipulations. This idea is derived from the concept of *hybrid task* [EMS90], in which a given task is solved by completing additional objectives (*e.g*, trajectory tracking). The paper [HIA98] illustrates the concept of *hybrid task* by achieving a contacting task in an unknown environment, while the robot is visually guided. The sensor-based control has the advantage of achieving a given task without fusing different sensing modalities or designing complex states machines. For instance, in [ENSHW14] a gripper is controlled by combining pre-touch sensors feedback and tactile sensors feedback, for a grasping task. In [DSLSV07], the authors introduce an approach centered on human-robot cooperation combining vision and force control, in order to achieve safe human-robot interaction in a given working space.

This short review of sensor-based control, through the principle of task functions and its usage in robotics, emphasizes the wide application field of this kind of approach. Such a vast literature cannot obviously be summed up in these two latter sections. Only the most essential aspects for the understanding of this manuscript have been introduced. We refer the interested reader to dedicated textbook addressing the sensor-based control in robotics such as [SELB91], [C+96] or [CH08]. At our end, we concentrate on the application of the sensor-based control in the field of robot audition, as detailed in the next section shaping the control framework of AS paradigm.

3.2 Theoretical framework

3.2.1 General principle of task function

The development of exteroceptive sensors, which are now for most of them affordable and embeddable, has stimulated their use on robots. Such sensors include cameras, microphones, tactile sensors, laser and force sensor for instance. In robotic context, the exteroceptive sensors give a physical measurement of the interaction with the surrounding environment. The typical approach consists in regulating the interaction through the information gathered from the sensor. Flexible and reactive control of the interaction is then aimed as well as precision, speed and low computation cost.

One way to control the robot is performed by open-loop techniques. The sensing of the environment is performed in a first step and the motion is applied to the robot afterwards. As illustrated by the Figure 3.2, the approach is based from an initial measurement $\mathbf{s}(\mathbf{r_0})$ obtained from the sensor at a position $\mathbf{r_0}$ and a desired pose $\mathbf{r}^*$. Therefore the approach is regulated by the control input $\dot{\mathbf{q}}$ of the robot until the desired pose, that corresponds to $\mathbf{r}(\mathbf{q}) = \mathbf{r}^*$ is reached from the extracted initial pose $\mathbf{r_0}(\mathbf{q})$, where $\mathbf{q}$ characterizes the configuration of the robot. This is typically the approach considered in sound source localization where the azimuth angle, the

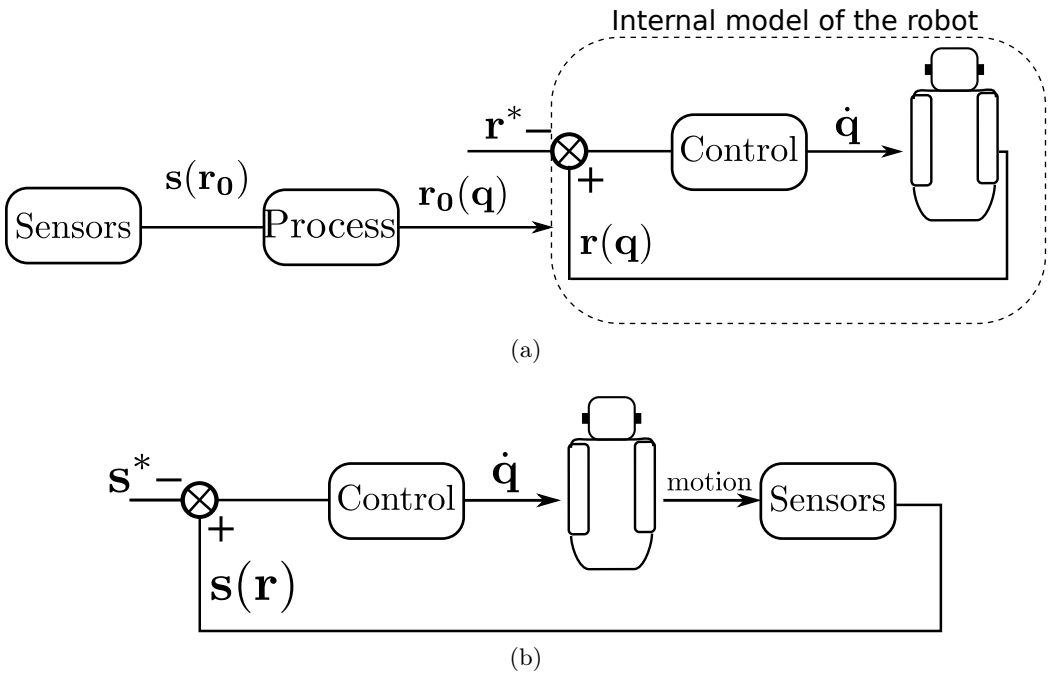


Figure 3.2: Two different approaches to control robots (a) An open-loop approach consists in deducing the pose of the robot $\mathbf{r_0}(\mathbf{q})$ with respect to a sensed target from a unique sensor measurement $\mathbf{s}(\mathbf{r_0})$ and move the robot accordingly through $\dot{\mathbf{q}}$; (b) A sensor-based approach links the motion of the robot to the sensor measurement $\mathbf{s}(\mathbf{r})$ in a feedback loop until the robot reaches a desired configuration characterized by a demanded measurement $\mathbf{s}^*$.

elevation angle, or the distance can be used to position the robot. However estimating the initial pose $\mathbf{r_0}(\mathbf{q})$ from an unique sensor measurement is not always trivial, as illustrated by the sound localization limitations, and generally requires a prior model of the environment. Moreover when $\mathbf{r_0}(\mathbf{q})$ is extracted, a motion planning strategy is then required to reach the desired pose from the actual pose of the robot. This action is perform "blindly" or more exactly in a "deaf manner" in our context, since $\mathbf{r}(\mathbf{q})$ is not related to the sensor information anymore.

By contrast the sensor-based control is designed by a closed-loop control that consists in a dynamic "sense and move" approach. The control input $\dot{\mathbf{q}}$ of the robot is steered by the dynamics feedback of the sensor measurements $\mathbf{s}(\mathbf{r})$ and generates the motion of the robot until a pose characterized by the sensors information $\mathbf{s}(\mathbf{r}) = \mathbf{s}^*(\mathbf{r}^*)$, is reached. Formally, the goal of the sensor-based task consists in controlling the system by regulating the error in the features between the observed and desired values $\mathbf{s}(\mathbf{r}) - \mathbf{s}^*(\mathbf{r}^*)$ to 0. By expressing the motion task at the sensory level, the step of estimating the desired pose of the robot is skipped (unless a PBVS approach is used). Likewise the motion planning is avoided. Hence this approach may be less conditioned by the knowledge of the environment, and may provide better solutions in unpredictable scenarios. The core principle to perform the latter regulation lies

in controlling the robot from a C^2-differentiable error function $\mathbf{e}(\mathbf{q}, t)$, that has to be regulated during a time interval $[0, T]$. This is the principle of the task function introduced in [SELB91] that depends on the robot configuration $\mathbf{q}(t)$ at a time t. The task function to be regulated can be expressed in terms of a measure f of the state of the robot as:

$$\mathbf{e}(\mathbf{q}(t), t) = f(\mathbf{q}(t), t) - f(\mathbf{q}^*(t)) \tag{3.1}$$

Generally the function f corresponds to a feature $\mathbf{s}$ that underlies spatial information about the robot configuration and that evolves accordingly to the motion applied to the robot. Since $\mathbf{e}(\mathbf{q}(t), t)$ is differentiable, the derivative of (3.1) that conditions the regulation of the task function is defined as

$$\dot{\mathbf{e}}(\mathbf{q}(t), t) = \frac{\partial \mathbf{e}(\mathbf{q}(t), t)}{\partial \mathbf{q}(t)} \dot{\mathbf{q}}(t) + \frac{\partial \mathbf{e}(\mathbf{q}(t), t)}{\partial t}. \tag{3.2}$$

In (3.2), two terms can be distinguished. The first one, involving the task Jacobian $\mathbf{J}_e = \frac{\partial \mathbf{e}(\mathbf{q},t)}{\partial \mathbf{q}}$ characterizes the evolution of the task function according to the dynamics of robot configuration space. The second term reflects all changes in the task function $\mathbf{e}$ with respect to time. More exactly, "disturbances" caused by the sensed target from which is inferred the state of the robot. From the control perspective, the task function $\mathbf{e}(\mathbf{q}, t)$ is said to be admissible, that is to say well-conditioned if it satisfies some specific properties such as [ECR92]:

- The regularity condition: $\mathbf{J}_e$ should not be singular (*i.e.*, invertible) $\forall t \in [0, T]$.
- There exists an unique ideal trajectory $\mathbf{q}_r(t)$ so that $\mathbf{e}(\mathbf{q}_r(t), t) = 0$ $\forall t \in [0, T]$ starting from in initial given condition $\mathbf{q}_r(0) = \mathbf{q}_0$.

In our case we are particularly interested by the configuration where the sensors are mounted on the robot, since we generally consider robots embedding microphones for instance. Then, the motions of the sensors are directly governed by the robot configuration, so that the features $\mathbf{s}(\mathbf{r}, t)$ measured from the sensors can be written as

$$\mathbf{s}(\mathbf{r}, t) = \mathbf{s}(\mathbf{q}(t), t) \tag{3.3}$$

$\mathbf{s}(\mathbf{r}, t)$ depends on the configuration of the robot at the current time $\mathbf{q}(t)$ but also on the potential evolution of the sensed target at t, Thus, the task function given in (3.1) can also be expressed as

$$\mathbf{e}(\mathbf{q}(t), t) = \mathbf{C}(\mathbf{s}(\mathbf{q}(t), t) - \mathbf{s}^*), \tag{3.4}$$

where $\mathbf{C}$ is a combination matrix that allows to match the dimension of $\mathbf{s}$ with the number of controlled degrees of freedom (DOF), when the features are redundant. Thus, $\mathbf{C}$ is used when the dimension of $\mathbf{s}$ is greater than the number of DOF of the robot [SAA91]. Nonetheless in the cases developed in this thesis, this configuration does not occur and $\mathbf{C} = \mathbb{I}$. Consider, for simplicity, the case of a regulation task in which a desired constant value $\mathbf{s}^*$ is specified. This value can be either computed or experimentally measured by manually moving the robot to the desired configuration.

Omitting the time dependence of the feature $\mathbf{s}$ and the robot configuration $\mathbf{q}$, from (3.4), (3.2) is equivalent to

$$\dot{\mathbf{s}} = \frac{\partial \mathbf{s}}{\partial \mathbf{q}}\dot{\mathbf{q}} + \frac{\partial \mathbf{s}}{\partial t}, \tag{3.5}$$

which characterizes the variation of the features $\mathbf{s}$ with respect to the motion applied to the robot. In this case the "disturbance" term exhibited in (3.2) corresponds to $\frac{\partial \mathbf{s}}{\partial t}$ that characterizes the potential motion of the sensed target, independently from the sensors. The first term of (3.5) $\mathbf{J_s} = \frac{\partial \mathbf{s}}{\partial \mathbf{q}}$ and $\mathbf{J_s} \in \mathbb{R}^{k\times m}$ is the Jacobian feature matrix that can be decomposed from the chain rule as

$$\frac{\partial \mathbf{s}}{\partial \mathbf{q}} = \frac{\partial \mathbf{s}}{\partial \mathbf{r}}\frac{\partial \mathbf{r}}{\partial \mathbf{q}}. \tag{3.6}$$

The latter equation contains the robot Jacobian $\mathbf{J}_r = \frac{\partial \mathbf{r}}{\partial q}$ that links the motion of the sensors to the motion of the robot. Since we consider embedded sensors, this Jacobian solely depends on the position of the sensors on the robot and therefore is assumed to be known. The robot Jacobian is multiplied by the interaction matrix $\mathbf{L}_s = \frac{\partial \mathbf{s}}{\partial r}$ that characterizes the chosen task through $\mathbf{s}$, by linking the dynamics of the selected feature to the motion of the sensors [CH06]. If we assume that the target is motionless and does not induce any change in values of the features $\mathbf{s}$ and no potential disturbances are acting on the system, the second term of (3.5) $\frac{\partial \mathbf{s}}{\partial t} = 0$. Hence, once $\mathbf{s}$ is selected, the relation, characterized by $\mathbf{L_s}$, between a set of features $\mathbf{s}$ and the sensor velocity $\mathbf{v_s}$ is given by

$$\dot{\mathbf{s}} = \mathbf{L_s}\mathbf{v_s} \tag{3.7}$$

in which $\mathbf{L_s} \in \mathbb{R}^{k\times n=6}$ is a matrix sized by k the number of measurements and the dimension of $\mathbf{v_s}$ (se_3). The sensor velocity is defined by $\mathbf{v_s} = (\boldsymbol{v_s}, \boldsymbol{\omega_s})$ where $\boldsymbol{v_s} = (v_x, v_y, v_z)$ denotes the spatial linear velocity and $\boldsymbol{\omega_s} = (\omega_x, \omega_y, \omega_z)$ the angular velocity. By extension, when using (3.4) and (3.7), we obtain the relationship between the sensor velocity and the time variation of the error:

$$\dot{\mathbf{e}} = \mathbf{L_e}\mathbf{v_s}, \tag{3.8}$$

where $\mathbf{L_e} = \mathbf{L_s}$. Then, designing a task regulation mainly consists in *i*) selecting $\mathbf{s}$, *ii*) computing or measuring $\mathbf{s}^*$ and *iii*) modelling the task Jacobian $\mathbf{J}_s$ (via the interaction matrix $\mathbf{L}_s$) while ensuring that the admissibility conditions of the task are not violated. Eventually, under the hypothesis of motionless target, the dynamics of the task function to be regulated is given by

$$\dot{\mathbf{e}}(\mathbf{q}, t) = \mathbf{J_s}\dot{\mathbf{q}}. \tag{3.9}$$

3.2.2 Control scheme

As reported previously, the control scheme of the sensor-based control can either be designed in order to consider sensor features or directly the pose of the robot in the loop. In our context, we are mainly interested in designing the control scheme

directly from the auditory features since sound source localization happens to be extremely complex. From then on, a simple control scheme can be designed with a purpose of exponential decoupled decrease of the error [ECR92]. In this case, the time variation of the expected error should follow

$$\dot{\mathbf{e}} = -\lambda \mathbf{e} \tag{3.10}$$

with $\lambda > 0$ a gain that tunes the time to convergence. Then, by combining (3.7) and (3.10) we obtain

$$\mathbf{v_s} = -\lambda \mathbf{L_s^+} \mathbf{e} \tag{3.11}$$

where $\mathbf{L_s^+} \in \mathbb{R}^{n \times k}$ is the Moore-Penrose pseudo-inverse of the interaction matrix. The pseudo-inverse is used when the inverse is not defined, that is when $n \neq k$ or $n = k$ and $|\mathbf{L_s}| = 0$. When $k < n$ and the rank of $\mathbf{L_s}$ is k, the Moore-Penrose pseudo-inverse is defined as

$$\mathbf{L_s^+} = \mathbf{L_s^\top} (\mathbf{L_s} \mathbf{L_s^\top})^{-1} \tag{3.12}$$

while when $k > n$, and the rank of $\mathbf{L_s}$ is n

$$\mathbf{L_s^+} = (\mathbf{L_s^\top} \mathbf{L_s})^{-1} \mathbf{L_s^\top}. \tag{3.13}$$

Otherwise, when $k = n$ and $|\mathbf{L_s}| \neq 0$ we simply have $\mathbf{L_s^+} = \mathbf{L_s^{-1}}$. The same reasoning can be developed by considering the feature Jacobian $\mathbf{J_s}$, when the sensors are embedded on a robot. Thus, the control input of the robot is given by

$$\dot{\mathbf{q}} = -\lambda \mathbf{J_s^+} \mathbf{e}. \tag{3.14}$$

In this case since $\mathbf{C} = \mathbb{I}$, we have $k \leq m$, and only the left pseudo-inverse given by (3.12) is used. Nonetheless in real configurations, it is impossible to know perfectly in practice either $\mathbf{L_s}$ or $\mathbf{L_s^+}$ (and by consequence $\mathbf{J_s}$ and $\mathbf{J_s^+}$). In the first case, the interaction matrix could depend on a quantity that cannot be directly measured by using the sensors. A typical example concerns the IBVS interaction matrix that contains the depth of the image feature relative to the camera frame that cannot be obtained from the image. Thus generally an approximation $\widehat{\mathbf{L_s}^+}$ of the inverse of the interaction matrix is used in the control loop. Consequently the control input of the system is more exactly defined as:

$$\mathbf{v_s} = -\lambda \widehat{\mathbf{L_s^+}} \mathbf{e} \tag{3.15}$$

Several ways exist to compute $\widehat{\mathbf{L_s}}$ and these different strategies affect the behaviour of the system and possibly the stability of the control scheme (see Section 3.2.3). Some of these strategies exposed in [CH06] are thoroughly detailed and studied in a practical use case in Chapter 5.3.2.2. Such type of regulation task allows the development of control methods by directly linking perception to action. To some extent, sensor-based control can be referred to a simplified modelling of human cognitive processes, that constantly update its behaviour from the sensory feedback. More precisely sensor-based control may be considered as a bio-mimicking process for

robot control and may underlie several advantages, as detailed in the previous section, especially when considering the hearing sense. The works exposed in Chapter 2.1.2 reported that the human auditory system is a dynamic process linking motion and perception. Until now this process has not been fully exploited in robot audition, while it is usually adopted with other sensory modalities.

3.2.3 Stability

In this part, we consider the fundamental issues related to the stability of the controllers. This analysis is important to guarantee that the solution $\mathbf{s}^*(\mathbf{r}^*)$ for which $\mathbf{e} = 0$ when $\mathbf{r}^*$ is reached, is an equilibrium point that attracts and can be reached by the control scheme. The stability thus ensures that the task can be achieved by the controller. More precisely, the controller is able to converge towards the configuration for which $\mathbf{e} = 0$.

To assess the stability of the closed-loop system, the Lyapunov analysis is generally used, since it is designed for non-linear dynamic systems. Conceptually the Lyapunov theory states that if a configuration $\mathbf{s}(\mathbf{r})$ that starts near the equilibrium point $\mathbf{s}^*(\mathbf{r}^*)$ stay near $\mathbf{s}^*(\mathbf{r}^*)$ forever, then the system is stable. Additionally, the asymptotic stability is obtained if the system is stable regarding the previous definition and all configurations $\mathbf{s}(\mathbf{r})$ that start near $\mathbf{s}^*(\mathbf{r}^*)$ converge towards $\mathbf{s}^*(\mathbf{r}^*)$.

The local asymptotic stability can be demonstrated from a stability criterion that makes use of a Lyapunov function candidate $\mathcal{L}$. This function should decrease along with the system state trajectories. For a system $\dot{x} = f(x)$ having a point of equilibrium at $x = 0$ the candidate function $\mathcal{L}(x) : \mathbb{R}^n \to \mathbb{R}$ ensure the local asymptotically stability under the following conditions:

$$\begin{aligned} &\bullet\ \mathcal{L}(x) = 0 \text{ if and only if } x = 0. \\ &\bullet\ \mathcal{L}(x) > 0\ \forall x \neq 0. \\ &\bullet\ \dot{\mathcal{L}}(x) < 0\ \forall x \neq 0. \end{aligned} \tag{3.16}$$

By extension, the global stability is obtained by showing that the control system is radially unbounded through

$$\|x\| \to \infty \Rightarrow \mathcal{L}(x) \to \infty. \tag{3.17}$$

The global asymptotic stability ensures that the system is attracted by the equilibrium point independently of the starting pose and is thus not restricted to the neighborhood of the equilibrium point.

For the specific sensor-based control exposed in this thesis, the system is defined by (3.8). We then consider the candidate Lyapunov function defined by the squared error norm $\mathcal{L}(\mathbf{e}) = \frac{1}{2}\|\mathbf{e}\|^2$. This function inherently satisfies the two first conditions of the local asymptotically stability (3.16). In order to fulfill the third condition we derive the candidate function as follows:

$$\dot{\mathcal{L}}(\mathbf{e}) = \mathbf{e}^\top \dot{\mathbf{e}}. \tag{3.18}$$

The latter equation can be detailed by injecting (3.8) and (3.11) in it:

$$\dot{\mathcal{L}}(\mathbf{e}) = -\lambda \mathbf{e}^\top \mathbf{L}_\mathbf{s} \widehat{\mathbf{L}_\mathbf{s}^+} \mathbf{e}. \tag{3.19}$$

As a consequence the global asymptotic stability of the control law (3.11), following Lyapunov definition, is ensured as soon as

$$\mathbf{L_s}\widehat{\mathbf{L_s^+}} > 0. \tag{3.20}$$

Yet, it is not always possible to ensure this condition, that consists in proving that the product in (3.20) is positive definite. A thorough analysis of the eigenvalues of the symmetric part of $\mathbf{L_s}\widehat{\mathbf{L_s^+}}$ would probably provide boundaries for the approximation of the interaction matrix $\widehat{\mathbf{L_s^+}}$. Such complex demonstration are not required for the comprehension of this thesis and we refer the interested reader to [MMR10] that consists in a typical study of the stability issue. Nonetheless, some essential stability results supporting AS framework are provided for different task configurations considered in the manuscript. In general, it is assessed that if the number of features is less or equal to the number of sensor DOF (*i.e.*, $k \leq n$), the global asymptotically stability condition (3.20) is ensured if the approximations involved in $\widehat{\mathbf{L_s^+}}$ are not too coarse [CH06].

3.2.4 Virtual linkages

Additionally to the stability, an important notion used in this thesis concerns the task that is realized by the robot. Introduced in [SELB91] and applied to VS [CRE93], virtual linkages provide a theoretical modelling of the task. Virtual linkages representation is a tool for analyzing the sensor-based task directly from the properties of the interaction matrix modelled. From the interaction matrix, it can be anticipated the behaviour of the robot regarding the completion of a task, that is to say the set of allowed poses for the robot to complete the given task. The concept is derived from the definition of mechanical linkages that characterize a mechanical structure of assembled and connected bodies or links. The connections between the links are modelled as providing movement of pure rotation or translation. Mechanical linkages are usually designed to transform a given input force and movement into a desired output force and movement. A simple illustration of mechanical linkage is materialized by robotic arms, where each link building up the arm is connected by a joint with a prismatic or rotoid motions.

In our context, a virtual link is defined between the robot and the sensed object (*i.e*, the sound source) and governs the admissible motion of the robot for completing a task. Let $\mathbf{v}^*_{\mathbf{s_i}}$ be a virtual motion of the sensor (or the robot), so that the sensor measure $\mathbf{s_i}$ remains constant. Thus from (3.7), $\mathbf{v}^*_{\mathbf{s_i}}$ is a solution of the equation

$$\mathbf{L}_{\mathbf{s_i}}\mathbf{v}^*_{\mathbf{s_i}} = 0 \tag{3.21}$$

and a skew reciprocal to $\mathbf{L}_{\mathbf{s_i}}$. The set of motions $\mathbf{v}^*_{\mathbf{s_i}}$ that leaves the set of features $\mathbf{s}$ unchanged is then denoted $\mathbf{S}^*$. $\mathbf{S}^*$ is a vector subspace characterizing these motions that is defined by

$$\mathbf{S}^* = \text{Ker } \mathbf{L}_s \tag{3.22}$$

where each column characterizes a motion of $\mathbf{v}^*_{\mathbf{s_i}}$ type. Thus it can be defined that a set of independent and compatible constraints satisfying $\mathbf{s}(\mathbf{r}) - \mathbf{s}^* = 0$ builds up

a virtual linkage between the sensor and the sensed object [SELB91]. It should be noticed that the vector subspace $\mathbf{S}^*$ allows us to characterize the virtual links since $\mathbf{s}(\mathbf{r}) - \mathbf{s}^* = 0 \ \forall t$ implies that $\dot{\mathbf{s}} = 0$. Considering m as the rank of the interaction matrix, N the rank of $\mathbf{S}^*$ is equal to $n - m$. This rank N defines the class of the virtual link. Consequently, the class of the virtual linkage corresponds to the number of DOF, in translation or rotation, that are not constrained to complete the task since m and N complement each other.

3.3 Conclusion

Sensor-based control has been researched since several decades and from different perspectives, as illustrated by the vast literature supporting this field. This framework is nowadays widely developed in connection with the emergence of exteroceptive sensors and allows reactive control of the robot to sensory stimulus at a relatively low computational cost. The vision and the touch senses are the common sensory feedback used in this framework. The well-known *visual servoing* paradigm, alone, illustrates the wide range of application of sensor-based control. Robots physical interactions (*e.g.*, grasping tasks) also benefit from the sensor-based approach through paradigm such as *proximity servoing* or *tactile servoing*, that are centered on the range and touch sensing. Such paradigms in these fields have led to impressive developments and prove its efficiency solving specific problems.

Nonetheless, surprisingly the sensor-based approach remains unused in robot audition, despite the convincing results obtained when using the touch sense modality or vision modality. However, it can be shown that expressing the problem of auditory perception through a task function can greatly simplify the interactions based on sound. The task function expresses a voluntary motion of the robot to interact with the sound. The auditory feedback is used to position the robot with respect to one or several sound sources. The resulting approach, AS, relies principally on the dynamics of the auditory cues to steer the motion of the robot. The dynamics of auditory cues appears to be more robust to realistic environments than their intrinsic values. AS is thus a dedicated control framework centered on hearing perception for robots unlike the classic sound localization framework that are adapted to machines in a broad sense.

The main tool building up AS is the closed-loop control ensued from the definition of the task function formalism. The resulting control scheme depends principally on the modelling of the task Jacobian, that characterizes the relationship between the motion of the robot and the dynamics of the auditory features. The task Jacobian models accurately the latter relationship while ensuring some given conditions. The first condition concerns the task admissibility that is generally ensured as soon as the task Jacobian is invertible, while the second one concerns the stability of the controller that guarantees the convergence of the system towards the global solution of a given task. The development of auditory control scheme using this type of formalism is the main thread of the two next chapters.

Chapter 4

ILD-based aural servo

The first modelling and applications of the AS framework are introduced in the following chapter. More specifically, this chapter concerns the development of sensor-based control framework based on the ILD. ILDs are analyzed from a binaural perspective, considering a set of two microphones in a free-field area. In this chapter and the following dealing with ITDs, the scattering effect of the head between the microphones is voluntarily omitted. The issues related to the scattering effect of head mounted systems are introduced later in Chapter 6.

In robot audition literature, surprisingly, when considering free-field conditions, the ILD is merely exploited compared to the prominence of methods based on ITDs. Only few works such as [BG05] or [CCW06] try to develop localization systems using this cue. The main challenge about ILD cues comes from the fact that they cannot be explicitly associated to an exploitable source location, when using two microphones. Actually, in planar scenes, iso-ILD lies on a circle on which the source may be located. From this information, it cannot be deduced an explicit source location. As a result, in [BG05], the authors consider four microphones in order to infer the sound location, while in [CCW06], the authors use the intersection between ILD and ITD cues to extract the source position. Furthermore, both of these studies point out the high sensitivity of ILDs towards reverberation. As a consequence, most free-field localization approaches are based on ITDs, when considering binaural sensors. On the contrary, in this chapter, we develop a control scheme, based on ILD measurements derived from acoustic properties of sound propagation in free field conditions. This contribution emphasizes the fact that AS paradigm opens perspectives on the utilization of auditory cues that are hardly exploitable in a localization paradigm. To integrate ILDs in AS, the core idea of the method presented in this chapter consists in linking the dynamics of ILDs to the motions of the robot through an interaction matrix.

For this purpose, a geometrical analysis is performed first, in order to establish the link between the ILD and the sound source location in Section 4.1.2. This analysis also draws the limits and constraints related to ILD cues.

From this preliminary analysis, in Section 4.2 the dynamics of the ILD is derived with respect to the motion of a pair of microphones, so that an interaction matrix related to this cue can be inferred. Hereafter this result is integrated into a control

scheme, that is analyzed with respect to its stability properties, and the positioning tasks that can be achieved. The proposed framework is eventually validated through several experiments in real world and uncontrolled environments.

In the last segment of this chapter in Section 4.3, we also integrate a distance cue in the control scheme, the absolute sound energy measured by the microphones, in order to overcome the limitations of ILD-based tasks related to reverberation and distant sources. Following the same process as in the previous section, the interaction matrix related to this cue is computed. A task analysis of the new control scheme is also provided. The control modelling is finally validated in various situations and acoustic conditions on a mobile robot.

4.1 ILD modelling

4.1.1 Scene configuration

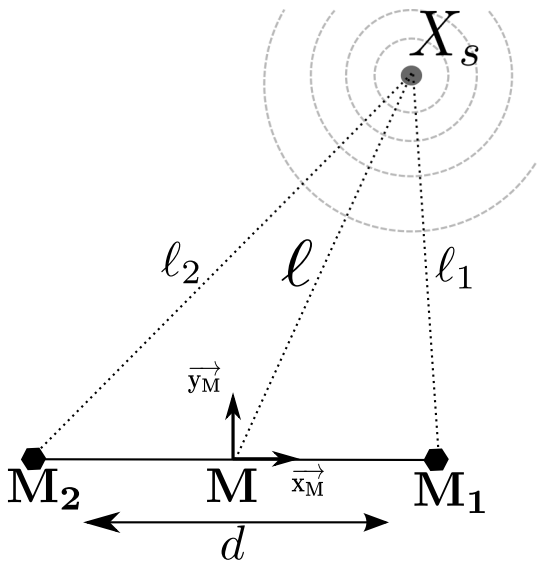


Figure 4.1: Geometric configuration of the considered system, that includes a source $\mathbf{X_s}$ emitting a spherical and uniform sound wave, and a pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$.

In a first phase, let us consider the ideal case of a motorized pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ in an area free of obstacle. The free-field microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ are separated by a distance d as illustrated in Figure 4.1. An omni-directional sound source $\mathbf{X_s}$ is continuously emitting a sound wave $a(t)$. For the following development $\mathbf{X_s}$ is shaped as a point that generates a sound wave uniformly in all directions. We also assume that the medium through which the sound travels is uniform. Furthermore, a frame $\mathcal{F}_m(\overrightarrow{x_M}, \overrightarrow{y_M})$ is attached to the midpoint of the microphones $\mathbf{M}$. In this frame, the Cartesian coordinates of each microphone are respectively $\mathbf{M_1}(\frac{d}{2}, 0)$ and $\mathbf{M_2}(-\frac{d}{2}, 0)$. The sound source $\mathbf{X_s}(x_s, y_s)$ is located at a distance ℓ_i from each

microphone $\mathbf{M_i}$. Considering the geometric modelling in Figure 4.1, the distances ℓ_i are respectively given by

$$\begin{cases} \ell_1 = \sqrt{(x_s - d/2)^2 + y_s^2} \\ \ell_2 = \sqrt{(x_s + d/2)^2 + y_s^2} \end{cases} \tag{4.1}$$

Eventually, let ℓ be the distance between $\mathbf{M}$, that is then immediately related to the source position through:

$$\ell = \sqrt{x_s^2 + y_s^2} \tag{4.2}$$

In this configuration, the microphones are endowed with 3 DOF in the horizontal plane, that are the translations along $\overrightarrow{x_M}$ and $\overrightarrow{y_M}$ axis, and the rotation around $\overrightarrow{z_M}$.

4.1.2 Geometrical properties of the ILD

Although the ILDs are mainly defined in the high frequencies of the spectral signal measured by the microphones, one can infer ILD directly from the acoustic properties of sound propagation in a similar process as [BG05]. Indeed, let us consider that $a(t)$, the signal generated by the sound source, is perceived by the microphone $\mathbf{M}_i$, for a frame of length w, under a high signal-to-noise ratio. In this condition, the signal perceived by each microphone $\mathbf{M}_i$ is geometrically defined as:

$$x_i(t) \propto \frac{a(t - \frac{\ell_i}{c})}{\ell_i}. \tag{4.3}$$

where $\frac{\ell_i}{c}$ expresses the sound propagation delay (see spherical sound wave propagation in Chapter 1.2.1). In the following, without loss of generality of our modelling we consider an unitary proportional gain. By integrating (4.3) over the frame of length w, the energy received by each microphone is defined as follows:

$$E_i = \int_{t=0}^{w} |x_i(t)|^2 \, \mathrm{d}t = \frac{1}{\ell_i^2} \int_{t=0}^{w} a^2 \left(t - \frac{\ell_i}{c} \right) \mathrm{d}t \tag{4.4}$$

Equation (4.4) characterizes the inverse-square law property inherent to spherical and uniform sound propagation as supposed in Section 4.1. The ILD ρ between the two microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ is then calculated from the ratio:

$$\rho = \frac{E_1}{E_2} = \frac{\ell_2^2 \int_{t=0}^{w} a^2 \left(t - \frac{\ell_1}{c} \right) \mathrm{d}t}{\ell_1^2 \int_{t=0}^{w} a^2 \left(t - \frac{\ell_2}{c} \right) \mathrm{d}t}. \tag{4.5}$$

Assuming that during w, the perceived sound signal is slowly varying, one can deduce that $\int_{t=0}^{w} a^2(t - \frac{\ell_1}{c}) \, \mathrm{d}t \approx \int_{t=0}^{w} a^2(t - \frac{\ell_2}{c}) \, \mathrm{d}t$. Actually, when the sound source is close to the microphones, we have $\frac{\ell_i}{c} \ll w$. And conversely, when the source is far, $\ell_1 \approx \ell_2$ since $\ell_i \gg d$. These observations are also consistent with ITDs that for realistic inter-microphones distances are generally lesser than 1 ms. In comparison, w has a length

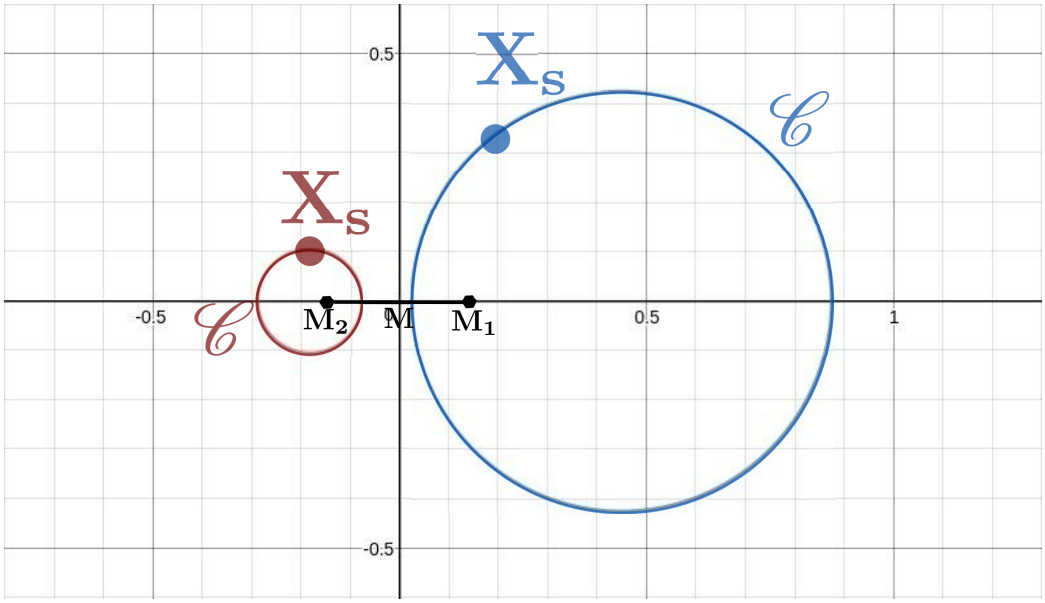


Figure 4.2: For a given sound source $\mathbf{X}_s$, the ILD refers geometrically to a circle $\mathscr{C}$, for which the *front-back ambiguity* should be solved.

of few tenths to hundreds ms, depending on the processing time that guarantees real-time interactions (*i.e.*, less than 150 ms). Consequently the ILD ρ can be simplified without significant loss of accuracy as:

$$\rho = \frac{\ell_2^2}{\ell_1^2} \tag{4.6}$$

Hence, from (4.6) and (4.5) the following relationship appears:

$$E_1 \ell_1^2 = E_2 \ell_2^2. \tag{4.7}$$

From these relationships, one can try to infer the sound location from the ILD measurement. With this aim in mind, by injecting the expression of each ℓ_i with respect to $\mathbf{X_s}$ in (4.7), we obtain

$$E_1 \left(x_s^2 - d x_s + \frac{d^2}{4} + y_s^2 \right) = E_2 \left(x_s^2 + d x_s + \frac{d^2}{4} + y_s^2 \right). \tag{4.8}$$

This relationship can also be expressed as

$$\left(x_s - \frac{d}{2} \frac{E1 + E2}{E1 - E2} \right)^2 + y_s^2 + \frac{d^2}{4} - \frac{d^2}{4} \frac{(E1 + E2)^2}{(E1 - E2)^2} = 0. \tag{4.9}$$

Finally (4.9) reduces to the equation of a circle $\mathscr{C}$ characterized by

$$(x_s - c_x)^2 + y_s^2 - \frac{E_1 E_2 d^2}{(E_1 - E_2)^2} = 0 \tag{4.10}$$

where $c_x = \frac{d}{2}\frac{E_1+E_2}{E_1-E_2}$. This result states that the sound source $\mathbf{X_s}$ is located "somewhere" on the circle $\mathscr{C}$, illustrated in Figure 4.2, centred on the point $(c_x, 0)$ with a radius of $c_r = d\left|\frac{\sqrt{E_1E_2}}{(E_1-E_2)}\right|$. In 3D scenes, the circle is changed into a sphere, that induces the *torus of confusion* when using both interaural cues (*i.e.*, ILD and ITD), as outlined in Chapter 2.1. Moreover the latter result implies that the circle $\mathscr{C}$ exists only if $E_1 \neq E_2$. Otherwise the circle degenerates into a straight line. Actually, when $E_1 = E_2$, with a similar manipulation as before (4.7) reduces to:

$$2dx_s = 0, \tag{4.11}$$

that corresponds to the bisection of the microphone pair.

At last, the position of the circle with respect to the microphones and interaural axis can be analyzed. Firstly, the circle $\mathscr{C}$ is always on the same lateral side of the microphones. Indeed $|c_x| \geq c_r$ since $|c_x| - c_r = \frac{d}{2|E_1-E_2|}(\sqrt{E_1} - \sqrt{E_2})^2$ which is a degree-two polynomial identity. Thus, for all sources $\mathbf{X_{s_i}}$ that belong to the same circle $\mathscr{C}$, $sign(x_{s_i})$ is the same. It can thus be said from ILD measurements if the source is on the left or right side of the microphones. But on the other hand, one cannot infer if the sound source is in the front or in the back side of the microphones since (4.10) imposes the circle $\mathscr{C}$ to be centered on a point $(c_x, c_y = 0)$. This observation can be related to the well-known *front-back ambiguity* although no exact azimuth direction nor distance can be extracted from the ILD.

4.2 A typical ILD-based interaction

4.2.1 ILD interaction matrix

For now on, this section is concerned about linking the motion of the robot to the ILD dynamic, so that an interaction matrix can be extracted. Indeed building the interaction matrix is one of the foundation of AS. In the configuration and analysis depicted above, one can relate the ILD dynamics to the motion of the microphones. From (4.6), the dynamics of the ILD can be obtained from the time variation of ρ that is given by:

$$\dot{\rho} = \frac{d}{dt}\left(\frac{\ell_2^2}{\ell_1^2}\right) = 2\frac{\ell_2\dot{\ell}_2\ell_1 - \dot{\ell}_1\ell_2^2}{\ell_1^3}, \tag{4.12}$$

that can be equally written as

$$\dot{\rho} = 2\left(\frac{\ell_2\dot{\ell}_2}{\ell_1^2} - \frac{\dot{\ell}_1}{\ell_1}\rho\right). \tag{4.13}$$

Consequently by replacing each ℓ_i by their value given in (4.1) one obtain:

$$\dot{\rho} = \frac{\dot{x_s}(2x_s + d) + 2y_s\dot{y_s}}{\ell_1^2} - \frac{\dot{x_s}(2x_s - d) + 2y_s\dot{y_s}}{\ell_1^2}\rho. \tag{4.14}$$

Equation (4.14) can also be expressed as a matrix relationship:

$$\dot{\rho} = \mathbf{L}_\rho\mathbf{v_M} \tag{4.15}$$

where $\mathbf{L}_\rho$ is the interaction matrix as introduced in Chapter 3.2.2, and $\mathbf{v_M} = (v_x, v_y, v_z, \omega_x, \omega_y, \omega_z)$ the spatial velocity of $\mathcal{F}_m$. From the basic kinematic equation [CH06]

$$\dot{\mathbf{X}}_\mathbf{s} = -\boldsymbol{v_s} - \boldsymbol{\omega_s} \times \mathbf{X_s} \Leftrightarrow \begin{cases} \dot{x}_s = -v_x - \omega_y z_s + \omega_z y_s \\ \dot{y}_s = -v_y - \omega_z x_s + \omega_x z_s \\ \dot{z}_s = -v_z - \omega_x y_s + \omega_y x_s \end{cases} \tag{4.16}$$

which relates the velocity of a 3-D point $\mathbf{X_s}$ to the sensor spatial velocity $\mathbf{v_s}$, (4.14) becomes

$$\dot{\rho} = v_x \frac{2x_s(\rho-1) - d(\rho+1)}{\ell_1^2} + v_y \frac{2y_s(\rho-1)}{\ell_1^2} + \omega_z \frac{y_s d(\rho+1)}{\ell_1^2}. \tag{4.17}$$

Eventually referring to (4.15), the interaction matrix related to ρ in the planar scene that we consider is identified as:

$$\mathbf{L}_\rho = \begin{bmatrix} \frac{2x_s(\rho-1)-d(\rho+1)}{\ell_1^2} & \frac{2y_s(\rho-1)}{\ell_1^2} & 0 & 0 & 0 & \frac{y_s d(\rho+1)}{\ell_1^2} \end{bmatrix}. \tag{4.18}$$

In connection with the geometric configuration, we consider only the non-zero terms of this matrix so that $k = 3$. In this planar scene, $\mathbf{v_M}$ is replaced by $\mathbf{u_M} = (v_x, v_y, \omega_z)$ so that $\mathbf{L}_\rho$ can be reduced to

$$\mathbf{L}_\rho = \begin{bmatrix} \frac{2x_s(\rho-1)-d(\rho+1)}{\ell_1^2} & \frac{2y_s(\rho-1)}{\ell_1^2} & \frac{y_s d(\rho+1)}{\ell_1^2} \end{bmatrix}. \tag{4.19}$$

For more consistence with the scene configuration and the reference frame $\mathcal{F}_m$, the interaction matrix $\mathbf{L}_\rho$ is redefined as a function of ℓ as

$$\mathbf{L}_\rho = \begin{bmatrix} \frac{2x_s(\rho-1)-d(\rho+1)}{\ell^2+\frac{d^2}{4}-dx_s} & \frac{2y_s(\rho-1)}{\ell^2+\frac{d^2}{4}-dx_s} & \frac{y_s d(\rho+1)}{\ell^2+\frac{d^2}{4}-dx_s} \end{bmatrix}, \tag{4.20}$$

from (4.1) that implies that $\ell_1^2 = \ell^2 + \frac{d^2}{4} - dx_s$. This matrix contains terms that are unknown, namely the source position x_s, y_s, and ℓ that is dependent on the two previous parameters as stated in (4.2). The approximated interaction matrix is then

$$\widehat{\mathbf{L}_\rho} = \begin{bmatrix} \frac{2\widehat{x}_s(\rho-1)-d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{2\widehat{y}_s(\rho-1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{\widehat{y}_s d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} \end{bmatrix}. \tag{4.21}$$

However, regarding the interaction matrix, the approximation of these parameters should be chosen carefully. Indeed, $\widehat{\mathbf{L}}_\rho$ has elements of infinite value as soon as $\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s = 0$. Thus $\widehat{y}_s$ and $\widehat{x}_s$ should be selected so that $\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s > 0$. Furthermore from (4.10) or (4.11) as soon as $\widehat{x}_s$ (respectively $\widehat{y}_s$) is set, it can immediately be deduced $\widehat{y}_s$ (respectively $\widehat{x}_s$) and then $\widehat{\ell} = \sqrt{\widehat{x}_s^2 + \widehat{y}_s^2}$. Nonetheless, the questioning about fixing the values of these parameters remains open. We address this issue in Section 4.2.4, by considering the Lyapunov stability of the controller as the matrix approximation criterion. As stated in the previous chapter, the Lyapunov global asymptotic stability of the control scheme ensures the convergence of the controller [CH06].

The latter interaction matrix (4.21) can also be expressed for ILD measured in dB. Indeed when considering the ILD ρ_{dB} defined as

$$\rho_{dB} = 10 \log_{10} \rho, \tag{4.22}$$

the time variation of the ILD expressed in dB is given by

$$\dot{\rho}_{dB} = \frac{10\dot{\rho}}{\rho \ln(10)}. \tag{4.23}$$

By injecting (4.15) into the latter equation, the interaction matrix related to ρ_{dB} is therefore defined as

$$\mathbf{L}_{\rho_{dB}} = \frac{10}{\rho \ln(10)} \mathbf{L}_{\rho}. \tag{4.24}$$

Eventually the approximated interaction matrix $\widehat{\mathbf{L}_{\rho_{dB}}}$ can be explicitly developed as

$$\widehat{\mathbf{L}_{\rho_{dB}}} = \frac{10}{\ln(10)} \begin{bmatrix} \frac{2\widehat{x}_s(10^{\frac{\rho_{dB}}{10}}-1)-d(10^{\frac{\rho_{dB}}{10}}+1)}{10^{\frac{\rho_{dB}}{10}}(\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s)} & \frac{2\widehat{y}_s(10^{\frac{\rho_{dB}}{10}}-1)}{10^{\frac{\rho_{dB}}{10}}(\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s)} & \frac{\widehat{y}_s d(10^{\frac{\rho_{dB}}{10}}+1)}{10^{\frac{\rho_{dB}}{10}}(\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s)} \end{bmatrix}. \tag{4.25}$$

It should be noticed that the expressions of $\widehat{\mathbf{L}_{\rho_{dB}}}$ in (4.25) and $\widehat{\mathbf{L}_{\rho}}$ in (4.21) are equivalent in terms of control. Depending on how the ILD is expressed one might choose the corresponding interaction matrix. In this thesis, we based our development on $\widehat{\mathbf{L}_{\rho}}$ that has a simpler form, and does not require additional calculation.

4.2.2 Control scheme

From the latter interaction matrix, we can then formulate a task that consists in positioning the robot so that given conditions characterized by the acoustic features are satisfied. The task is performed by considering a single ILD measurement $\rho(t)$ extracted from the sound signal and by minimizing the error $\|e(t)\|$ in the same process described in Chapter 3.2.1 . Hence the task is characterized by an error

$$e(t) = \rho(t) - \rho^* \tag{4.26}$$

where ρ^* denotes the measurements for the desired ILD value. From (4.15) and with a perfectly known interaction matrix as (4.20), it is possible to design a control scheme with a purpose of exponential decoupled decrease of the error [SELB91] in which the error follows

$$\dot{e} = -\lambda e, \tag{4.27}$$

where $\lambda > 0$ is a gain that tunes the time to convergence. As stated in the previous section the actual matrix $\mathbf{L}_{\rho}$ is not fully known since some of its elements depend on the sound source position. Hence we design a control scheme governed by the approximation of the interaction matrix $\widehat{\mathbf{L}_{\rho}}$ in (4.21), in which the velocity of the microphones is computed from (4.15) [CH06] as

$$\mathbf{u_M} = -\lambda \widehat{\mathbf{L}_{\rho}^{+}} e. \tag{4.28}$$

In this latter equation $\widehat{\mathbf{L}_\rho^+} \in \mathbb{R}^{3\times 1}$ is the Moore-Penrose pseudo-inverse of $\widehat{\mathbf{L}_\rho}$ (see Chapter 3.2.1). Thus the control scheme can be explicitly written as

$$\mathbf{u_M} = -\lambda \frac{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s}{\widehat{N_1}^2 + \widehat{N_2}^2 + \widehat{N_3}^2} \begin{bmatrix} 2\widehat{x}_s(\rho-1) - d(\rho+1) \\ 2\widehat{y}_s(\rho-1) \\ \widehat{y}_s d(\rho+1) \end{bmatrix} (\rho - \rho^*), \tag{4.29}$$

where

$$\begin{cases} \widehat{N_1} = 2\widehat{x}_s(\rho-1) - d(\rho+1) \\ \widehat{N_2} = 2\widehat{y}_s(\rho-1) \\ \widehat{N_3} = \widehat{y}_s d(\rho+1) \end{cases} \tag{4.30}$$

4.2.3 Task analysis

From the interaction matrix in (4.21), it is then possible to achieve positioning tasks with the control scheme (4.29). However it is necessary to analyze and understand what is the task performed through the control scheme developed in the previous section. As introduced in Chapter 3.2.4, we make use of virtual linkages in order to exhibit the motions that should be expected from the control scheme. In this approach, a vector subspace $\mathbf{S}^*$ is defined so that it represents all motions of the sensors for which the ILD ρ remains constant. To some extent this approach can be seen as similar the development produces in Section 4.1.2, in which $\mathscr{C}$ characterizes the motions of the source that gives the same ILD measurement. The subspace $\mathbf{S}^*$ is defined more explicitly as:

$$\mathbf{S}^* = \text{Ker } \mathbf{L}_\rho. \tag{4.31}$$

Hence the subspace is defined as

$$\mathbf{S}^* = \begin{bmatrix} \mathbf{u}^*_{\mathbf{M_1}} & \mathbf{u}^*_{\mathbf{M_2}} \end{bmatrix} = \begin{bmatrix} v_{x1} & v_{x2} \\ v_{y1} & v_{y2} \\ \omega_{z1} & \omega_{z2} \end{bmatrix}. \tag{4.32}$$

When considering the interaction matrix $\mathbf{L}_\rho$ given in (4.20), we obtain:

$$\mathbf{S}^* = \begin{bmatrix} y_s d(1+\rho) & 2y_s(\rho-1) \\ 0 & d(\rho+1) + 2x_s(1-\rho) \\ d(\rho+1) + 2x_s(1-\rho) & 0 \end{bmatrix}. \tag{4.33}$$

Although it is not trivial to infer the geometrical motions that come out from (4.31), they can be deduced by geometric construction. The first vector $\mathbf{v}^*_{\mathbf{M_1}}$ refers to a circular motion of the microphones for which $\rho = \frac{\ell_2^2}{\ell_1^2}$ remains constant. Thence it can be immediately deduced that this vector refers to a circular motion around the source, in which ℓ_1 and ℓ_2 remain constant as depicted in Figure 4.3. Concretely, from a given orientation of the microphones with respect to the source, moving around the sound source by keeping the same relative orientation does not change the ILD value. Indeed in this configuration the circles $\mathscr{C}$ all intersect on the actual source position.

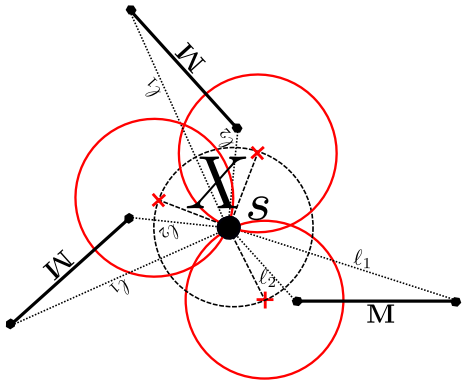
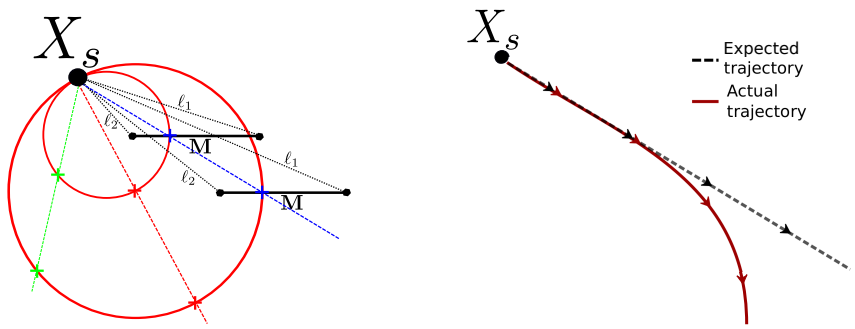


Figure 4.3: Admissible poses of the microphones for a given ρ

As for, the second vector $\mathbf{v}^*_{\mathbf{M_2}}$ refers to a translation. Any translation of the microphones directly induces variation of ℓ_1 and ℓ_2. As a consequence, for maintaining the ILD unchanged, the constraint $\rho = \frac{k\ell_2^2}{k\ell_1^2}$ rises, with $k \in \mathbb{R}$. Geometrically, the translation of the microphones implies that each point belonging to $\mathscr{C}$ is moved in the direction of the sound source. Therefore for each position with the same ratio ρ, we obtain circles $\mathscr{C}$ that are tangent to the actual position of the sound source as illustrated in Figure 4.4a. However, it should be denoted that both $\mathbf{v}^*_{\mathbf{M_1}}$ and $\mathbf{v}^*_{\mathbf{M_2}}$ depends on the ratio ρ that has a limited range (or resolution) in the far-field (see Section 4.2.6), which may modify these motions. As illustration Figure 4.4b shows the motion expected from $\mathbf{v}^*_{\mathbf{M_2}}$ and the actual motion obtained because of the properties of the ILD in the far-field.



(a) (b)

Figure 4.4: Translation motion of the microphones for a given ρ in (a). However the motion degenerates in the far-field because of the limitations of the ILD discussed in Section 4.2.6

Eventually, by considering ideal conditions (*i.e* not limited ILDs), any linear combination of these two motions is possible which implies that infinite poses exist to complete a given task governed by the control scheme. These poses can be represented by concentric circles of radius c_r around the sound source. Hence, a typical task involving the use of the ILD mainly consists in orienting the robot towards a

given direction with respect to the source location. This result implies that controlling only one DOF, namely the rotation ω_z is enough to complete any task involving the regulation of the ILD. Hence, the task defined by the ILD in AS is consistent with the exploitation of ILD cues for human listeners, since interaural cues are physiologically related to an azimuth direction as exposed in Chapter 2.1.2. However as stated Section 4.1.2, this cue does not explicitly provide an azimuth direction (*i.e.*, circle $\mathscr{C}$), which makes it complex to exploit for sound source localization paradigm, while AS can still benefit from this information.

4.2.4 Stability analysis

In this part, we study the stability conditions of the control scheme (4.29) with respect to the approximated interaction matrix $\widehat{\mathbf{L}_\rho}$. The stability properties tailor the approximation of the unknown parameters in the interaction matrix $\widehat{\mathbf{L}}_\rho$, since they condition the convergence of the control scheme (see Chapter 3.2.2). Since $\widehat{\mathbf{L}_\rho} \in \mathbb{R}^{1\times 3}$, the control scheme (4.29) is globally asymptotically stable when

$$\mathbf{L}_\rho \widehat{\mathbf{L}_\rho^+} > 0. \tag{4.34}$$

In a first step, in order to ease the comprehension and the reading of this analysis, let us redefine the interaction matrix $\widehat{\mathbf{L}}_\rho$ as

$$\widehat{\mathbf{L}}_\rho = \left[\frac{\widehat{N_1}}{(\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s)} \quad \frac{\widehat{N_2}}{(\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s)} \quad \frac{\widehat{N_3}}{(\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s)}, \right] \tag{4.35}$$

where $\widehat{N_1}$, $\widehat{N_2}$, $\widehat{N_3}$ are defined from (4.30). Similarly $\mathbf{L}_\rho$ is defined with N_1, N_2, N_3. From this convention of writing, (4.34) becomes

$$\mathbf{L}_\rho \widehat{\mathbf{L}_\rho^+} = \frac{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s}{\ell^2 + \frac{d^2}{4} - dx_s} \frac{N_1\widehat{N_1} + N_2\widehat{N_2} + N_3\widehat{N_3}}{\widehat{N_1}^2 + \widehat{N_2}^2 + \widehat{N_3}^2} > 0. \tag{4.36}$$

Knowing that $\widehat{\ell_1}^2 > 0$ or equivalently $\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s > 0$, (4.36) is true if and only if $N_1\widehat{N_1} + N_2\widehat{N_2} + N_3\widehat{N_3}$ is greater than 0. Thus one can rewrite the condition by developing the latter expression as:

$$4y_s\widehat{y}_s(\rho-1)^2 + y_s\widehat{y}_s d^2(\rho+1)^2 + (2x_s(\rho-1) - d(\rho+1))(2\widehat{x}_s(\rho-1) - d(\rho+1)) > 0 \tag{4.37}$$

It is then clear that when $E_1 = E_2$ (*i.e.*, $x_s = 0$ and $\rho = 1$), the stability is guaranteed as soon as $sign(y_s) = sign(\widehat{y}_s)$. However when $\rho \neq 1$, some supplementary analysis are required. To do so, let us rearrange the condition (4.37) to :

$$(\rho-1)^2(4y_s\widehat{y}_s + 4x_s\widehat{x}_s) + (\rho+1)^2 d^2(1 + y_s\widehat{y}_s) - (\rho+1)(\rho-1)2d(x_s + \widehat{x}_s) > 0. \tag{4.38}$$

By dividing (4.38) by $(\rho+1)^2$, we obtain a quadratic function $f(Z)$ depending on the ratio ρ that is

$$f(Z) = Z^2(4y_s\widehat{y}_s + 4x_s\widehat{x}_s) - 2Zd(x_s + \widehat{x}_s) + d^2(1 + y_s\widehat{y}_s) > 0, \tag{4.39}$$

with $Z = \frac{\rho-1}{\rho+1}$ and $Z \in]-1,1[$. It can be noticed that $sign(Z) = sign(x_s)$ from the geometrical properties outlined in Section 4.1.2. Indeed $x_s > 0$ when $\rho > 1$, and $x_s < 0$ when $\rho < 1$. Therefore, for a random variable $b \in]-1,1[$, $f(Z)$ is characterized by

$$\begin{cases} f(-b) = b^2(4y_s\widehat{y}_s + 4x_s\widehat{x}_s) + 2db(x_s + \widehat{x}_s) + d^2(1 + y_s\widehat{y}_s) \\ f(b) = b^2(4y_s\widehat{y}_s + 4x_s\widehat{x}_s) - 2db(x_s + \widehat{x}_s) + d^2(1 + y_s\widehat{y}_s) \end{cases} \tag{4.40}$$

From the relationship between the sign of x_s and the sign of Z, one can notice that $f(b) = f(-b)$ when $sign(\widehat{x}_s) = sign(x_s)$. Under this condition, knowing that $f(Z)$ is a parabola from its quadratic nature, the vertex of $f(Z)$ is the point at $(0, f(0))$. As a consequence, to fulfill the condition (4.39), it is sufficient to demonstrate that (i), $f(Z)$ is a parabola that opens upward, and (ii), the vertex value $f(0)$, which is a minimum when (i) is ensured, is greater than 0.

(i). First, it can be demonstrated that $f(Z)$ opens upward only if

$$(4y_s\widehat{y}_s + 4x_s\widehat{x}_s) > 0. \tag{4.41}$$

A sufficient condition to fulfill (4.41) consists in setting $sign(\widehat{y}_s) = sign(y_s)$ and $sign(\widehat{x}_s) = sign(x_s)$, that has already been assumed.

(ii). In this context, the vertex value $f(0)$ is greater than 0 when

$$d^2(1 + y_s\widehat{y}_s) > 0. \tag{4.42}$$

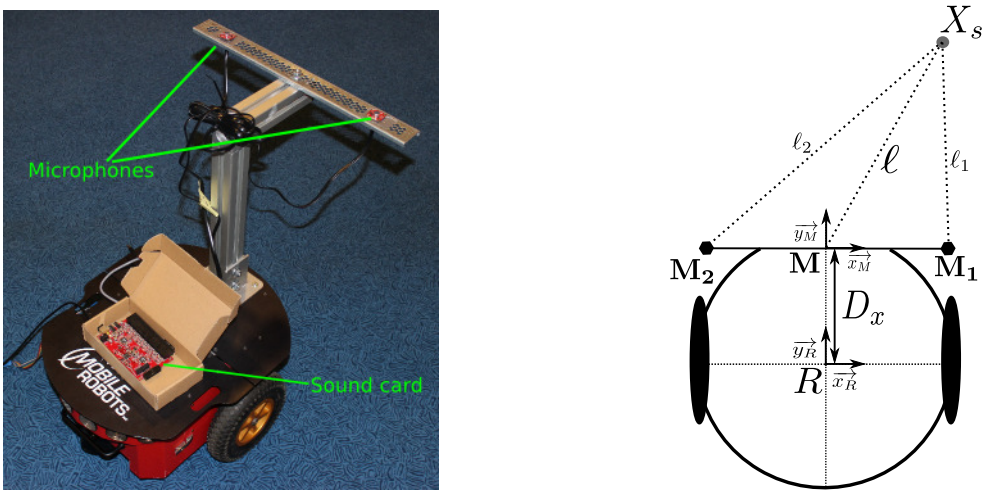
The latter condition is immediately completed when $sign(\widehat{y}_s) = sign(y_s)$.

Consequently the Lyapunov stability condition of $\widehat{\mathbf{L}_\rho}$ expressed through (4.39) is ensured as soon as $sign(\widehat{y}_s) = sign(y_s)$ and $sign(\widehat{x}_s) = sign(x_s)$. One can underline that $sign(\widehat{x}_s) = sign(x_s)$ is obtained directly by the measurement of the ILD ρ. Actually $sign(x_s) = sign(\rho - 1)$. However $sign(\widehat{y}_s) = sign(y_s)$ cannot be ensured from the ILD measurements because of the properties of the circle $\mathscr{C}$ depicted in Section 4.1.2. The latter condition directly refers to the *front/back ambiguity* that is unfortunately inherent to the binaural cues. Despite the *front/back ambiguity*, these excellent stability conditions confirm that the exact sound source location is not required for the controller. The stability conditions are not restrictive at all and let us foresee an infinite convergence domain.

In the context depicted until now, we assumed ideal acoustic conditions in free-field, that simplify our modelling but that are unfortunately not realistic, on the other hand. Indeed as discussed in Chapter 1.2 and in Chapter 2.3, one of main issue of robot audition is concerned about the consistency and the modelling of the auditory features in perturbed environment. Nevertheless, this study emphasizes the potential robustness of the controller, that is tolerant to rough approximations of the interaction matrix. This interesting property allows us to envisage that the control scheme may also be robust to approximate acoustic modelling, and by extension robust to realistic acoustic environment. This hypothesis is verified in the next section, that is articulated around the experimental validations of our approach in real world conditions.

4.2.5 Experimental validations

4.2.5.1 Preliminaries: robot modelling and control scheme



(a) (b)

Figure 4.5: Modelling of the robotic platform

For the experiments, we consider a non-holonomic unicycle robot *Pioneer 3DX* endowed with two omnidirectional microphones as illustrated on Figure 4.5. Schematically, the system consists in a planar robot controlled in the horizontal plane with the two microphones $\mathbf{M_1}$ and $\mathbf{M_2}$. In addition to $\mathcal{F}_m$ attached to the microphones, we define $\mathcal{F}_r(\overrightarrow{x_R}, \overrightarrow{y_R})$ attached to the robot. D_x denotes the distance between the center of the robot $\mathbf{R}$ and the midpoint of the microphones $\mathbf{M}$. The robot can be controlled upon two DOF: the control input $\dot{\mathbf{q}}$ is given by (u, ω), respectively the translation velocity along $\overrightarrow{y_R}$ and the angular velocity around $\overrightarrow{z_R}$.

Consequently the relationship between the dynamics of the ILD ρ and the control input $\dot{\mathbf{q}}$ is:

$$\dot{\rho} = \mathbf{J}_\rho \dot{\mathbf{q}} \tag{4.43}$$

where $\mathbf{J}_\rho$ is the Jacobian feature matrix introduced in Chapter 3.2.1 that is equal

$$\mathbf{J}_\rho = \mathbf{L}_\rho \mathbf{J_r}. \tag{4.44}$$

$\mathbf{J}_r$ being the robot Jacobian. From the modelling given in Figure 4.5, the robot Jacobian is equal to

$$\mathbf{J_r} = \begin{bmatrix} 0 & D_x \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{4.45}$$

Hence, the control scheme of the robot is (see Chapter 3)

$$\dot{\mathbf{q}} = -\lambda \widehat{\mathbf{J}_\rho^+} (\rho - \rho^*). \tag{4.46}$$

Furthermore, from the previous task analysis, we demonstrated that a task considering the ILD ρ mainly specifies a desired orientation of the robot w.r.t the sound source. Therefore a control input characterized by the angular velocity ω only is sufficient to achieve any task involving one sound source. In the following, we consider this particular case where u would always be equal to 0. Hence the Jacobian matrix $\widehat{\mathbf{J}_\rho}$ reduces to

$$\widehat{\mathbf{J}_\rho} = \frac{D_x(2\widehat{x}_s(\rho-1) - d(\rho+1)) + \widehat{y}_s d(\rho+1)}{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s}, \tag{4.47}$$

and the control input is then

$$\dot{\mathbf{q}} = -\lambda \frac{\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s}{D_x(2\widehat{x}_s(\rho-1) - d(\rho+1)) + \widehat{y}_s d(\rho+1)}(\rho - \rho^*). \tag{4.48}$$

4.2.5.2 Experimental results

The sets of experiments conducted in this section and the following are performed in a room with a reverberation time $\mathrm{RT}_{60} \approx 580$ ms. The microphones instrumenting the robot were connected to a sound card 8SoundsUSB [AC+] that processes the signal in real time. The sound card operates at a sampling frequency of 48 kHz, and provides frames of 256 samples. The sound energy is computed from 10 consecutive frames w (*i.e.*, 50 ms). The global processing time of one iteration corresponds approximately to the length of each frame. The parameters given in Table 4.1 were used for all experiments. In a first step, we do not account for the *front-back ambiguity*, and thus assume that the sound source is in the front side of the robot (e.g. $y_s > 0$).

d	0.31 m
D_x	0.3 m
$\widehat{y}_s$	1 m
$\widehat{x}_s$	sign$(\rho - 1)$ $\times 1$ m
λ	0.5

Table 4.1: Experimental settings

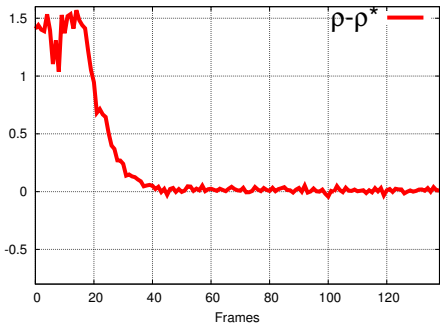
The first experiment simply consists in facing the sound source from a random orientation. The desired ILD is thus set to $\rho^* = 1$. The sound source is a loudspeaker emitting a white Gaussian noise. The SNR at the desired pose is around 20 dB in presence of typical noise such as computer noise and ventilation in the room. As illustrated in Figure 4.6, the control scheme allows to face the sound source starting from a random orientation of the robot. The magnitude of the error follows an exponential decrease behaviour as expected from the modelling of the control scheme. It should also be denoted that our approach is straightforward: the control input computed from (4.48) simply consists in a scalar corresponding to the angular velocity of the robot. Furthermore, our approach is performed without signal enhancement, modelling of the environment nor tracking, which explains the low computation cost of the control system.

(a)

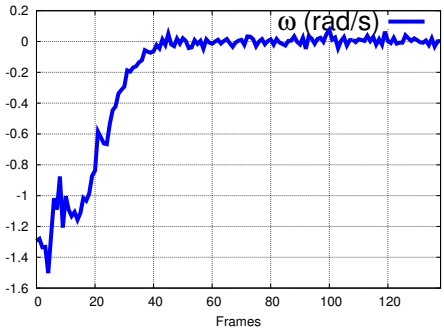
(b)

(c)

(d)



(e) Features error



(f) Velocities input

Figure 4.6: A typical task using the ILD ρ consists in facing the sound source from a random orientation

4.2.5.3 Addressing front-back ambiguity

Although the previous experiment confirmed the consistency of our approach, we assumed that the sound source was in the front side of the robot, which can be restrictive. The *front-back ambiguity* remains an issue that cannot be solved from the binaural cues values. The development in Chapter 2.1.2 showed that human

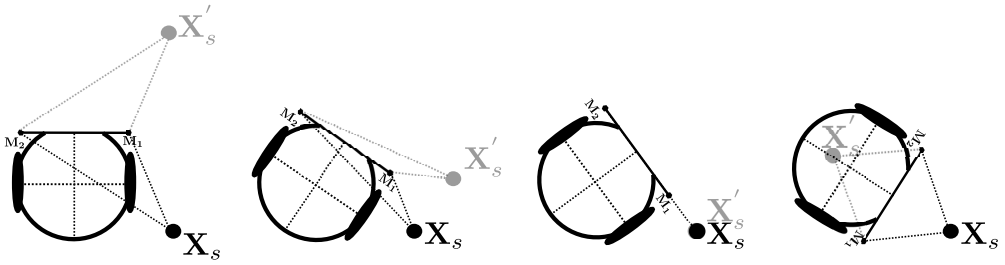


Figure 4.7: From left to right a position task that consists to face the sound source $\mathbf{X}_s$ located on the back-side of the robot: because of the front-back ambiguity the robot turns with respect to the phantom source $\mathbf{X}_s^{'}$. This behaviour naturally allows to complete successfully the task, since the motion of the robot will replace the actual source on the front-side

listeners disambiguate the location of the source from monaural cues (*i.e.*, action of the pinna through the HRTF) and from head motion. The first solution, based on monaural cues, cannot be applied to our context since the robot is endowed with free-field microphones, that is to say no HRTF is available. Yet, the motion of the robot can be used in order to dissipate the ambiguity. A simple method would consist in a state-machine inferring the side of the source from the evolution of the ILD with respect to the motion of the robot. This kind of approach has been applied in [MMWB15, MMB15] for instance. More sophisticated techniques have also been proposed in the literature such as stochastic filtering in [PDA12] or [NCVC16] that respectively use an unscented Kalman filter and mixture of Kalman filter. These two approaches fuse the motion of the robot to the cues measurements to infer the sound location without ambiguity.

Our approach is designed for situations where the sound source is in the front side (*i.e.*, $y_s > 0$) which explains that the interaction matrix is parameterized with $\widehat{y}_s > 0$ so that the stability conditions are ensured. Hence, the control scheme might not be stable when the sound source is located on the back-side since $sign(\widehat{y}_s) \neq sign(y_s)$, and the robot might not converge towards the desired pose. Under this hypothesis, the robot moves with respect to a phantom sound source (symmetric to the actual source) as illustrated in Figure 4.7. The robot moves towards an opposite direction to the one required to the completion of the task which increases the magnitude of the error $e = \rho - \rho^*$. This result is consistent with the stability conditions formulated in Section 4.2.4 that are fulfilled when $sign(\widehat{y}_s) = sign(y_s)$. This characteristic that could be seen as a limitation of the control scheme, turns out to be actually an advantage. Indeed, this characteristic induces that the only way to complete a given task necessitates $sign(\widehat{y}_s) = sign(y_s)$, that is to say the source to be located in the front side of the robot since we fixed $\widehat{y}_s > 0$. From the property of ILD-based task that consists in turning the robot towards a given direction with respect to the source until the condition $e = \rho - \rho^* = 0$ is met, it can be inferred that the latter condition cannot be fulfilled when the source is located in the back since the magnitude of the error increases. Consequently, the rotation motion will necessarily

lead to $y_s > 0$. Concretely the magnitude of the error increases in a first phase, then decreases towards zero as soon as $y_s > 0$ which allows completing the task. Thus, the robot always reaches a pose that corresponds to the expected configuration $y_s > 0$, in which the task can be completed similarly to the previous experiment. The *front-back ambiguity* is then inherently solved by our control modelling: without any assumption about the location of the source, it is possible to complete positioning tasks based on ILD measurements.

This result has been confirmed by experiments in which we observe the predicted behaviour discussed above. The experiment illustrated by Figure 4.8, shows a first phase where the error of the task increases until $y_s > 0$. Subsequently, from this configuration, an exponential decrease of the magnitude of error can be observed until the robot faces the sound source.

4.2.5.4 Addressing the case of a moving sound source

Eventually, we can tackle the problem related to a moving sound source. Considering robot audition state-of-the-art, dealing with dynamic source is a bottleneck of sound source localization, since it requires a tracking and modelling of the motions of the source, beside perturbation caused by the environment or the robot. By contrast, AS is a closed-loop control, that naturally induces flexibility and reactivity to any modification of the acoustic perception, thanks to the real-time feedback. Hence, any motion of the sound source will have an impact on the control input of the robot. Indeed when the source moves laterally, the magnitude of the error $e = \rho - \rho^*$ increases, which immediately generates a motion in order to reduce as fast as possible the error. Hence, this type of control scheme is particularly suitable for tracking a dynamic source. The experiment carried in Figure 4.9 sum up all the potential of AS. Initially, the source is in the back of the robot. This issue is correctly solved by the control scheme, and the robot is able to face the sound source. Afterwards, the sound source is moved laterally and arbitrary while being accurately tracked by the robot in real time. The motion of the robot maintains the source in the *auditory fovea*. However, fast motions of the sound source, may entails some delay in the control scheme to compensate the error. Nevertheless this error could be reduced by advanced control techniques. If we recall the principle of task function given in Chapter 3.2.1, the time derivative of ρ is exactly defined as

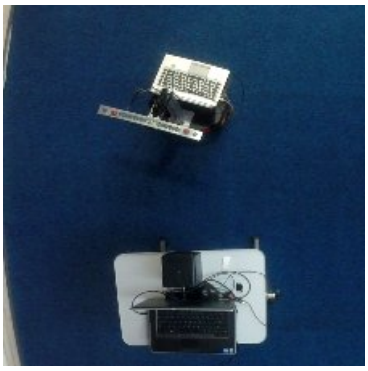
$$\dot{\rho} = \dot{e} = \mathbf{L_e v_M} + \frac{\partial e}{\partial t}, \tag{4.49}$$

in which $\frac{\partial e}{\partial t}$ characterizes the time variation of e due to the unknown source motion. Until now, to derive the control scheme from the dynamics of the ILD, we assumed motionless source, considering $\frac{\partial e}{\partial t} = 0$. Thus, in order to avoid this tracking error, $\frac{\partial e}{\partial t}$ should be integrated in the control scheme. With a purpose of exponential decoupled decrease of the error ($\dot{e} = -\lambda e$), the velocity of the microphones is therefore defined as

$$\mathbf{v_M} = -\lambda \widehat{\mathbf{L}_e^+} e - \widehat{\mathbf{L}_e^+} \widehat{\frac{\partial e}{\partial t}}, \tag{4.50}$$

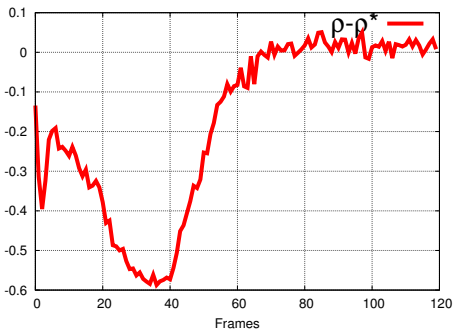
(a)

(b)

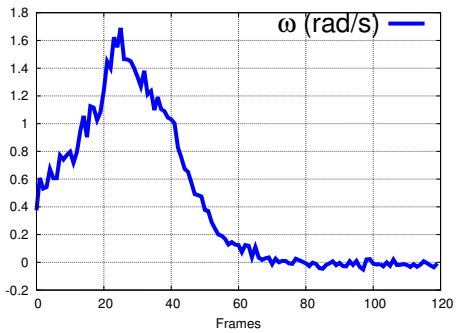


(c)

(d)



(e) Features error

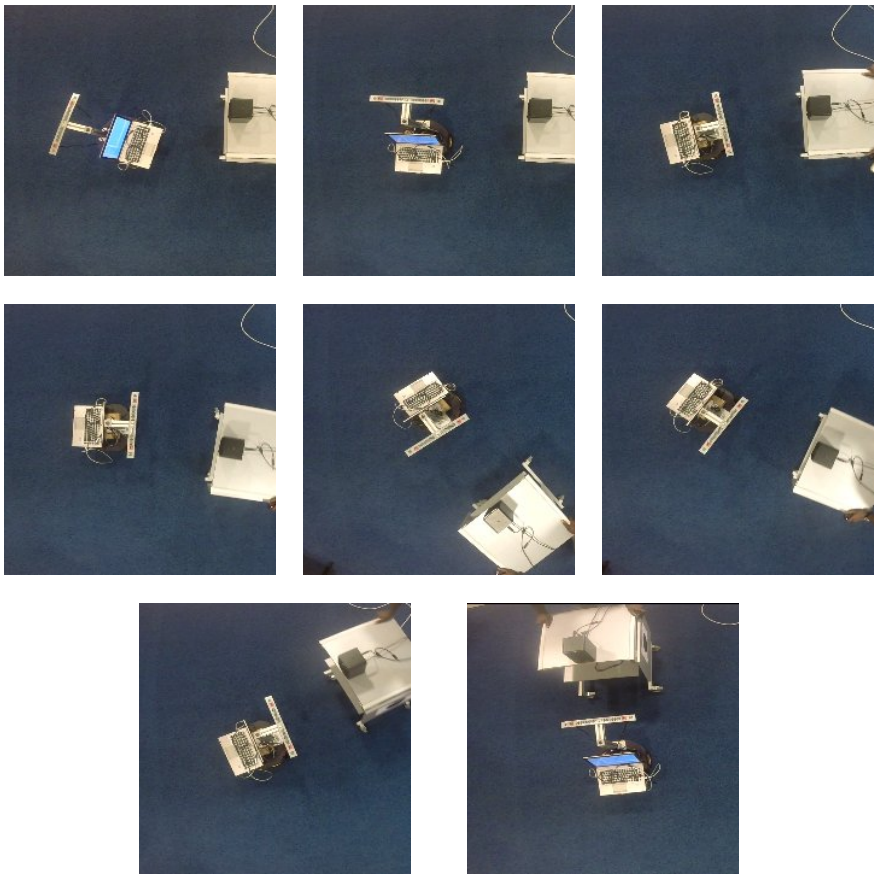


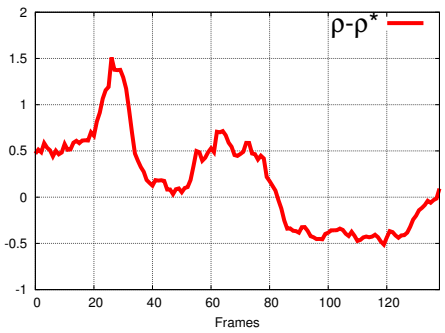
(f) Velocities input

Figure 4.8: The *front-back ambiguity* is inherently solved by the control scheme: In a first step the error increases until the configuration corresponds to the correct modelling (*i.e.*, $y_s > 0$). And in a second phase the robot converges towards a correct pose with an exponential decrease of the magnitude of the error.

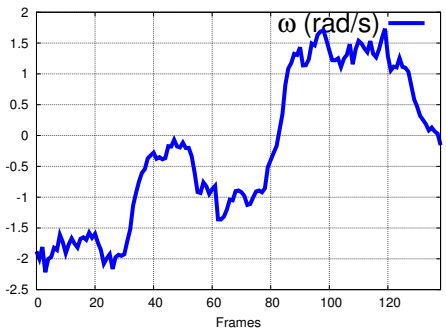
in which $\widehat{\frac{\partial e}{\partial t}}$ is an estimation of the time variation of e due to the source motion. Eventually the dynamics of the error is obtained by injecting (4.50) into (4.49) as:

$$\dot{e} = -\lambda \mathbf{L_e} \widehat{\mathbf{L}_e^+} e - \mathbf{L_e} \widehat{\mathbf{L}_e^+} \widehat{\frac{\partial e}{\partial t}} + \frac{\partial e}{\partial t}. \tag{4.51}$$





(i) Features error



(j) Velocities input

Figure 4.9: An experiment summing up all configurations (reading from left to right, top to bottom): first the front-back ambiguity is solved, in a second phase, the robot converges towards a satisfying pose, and finally the speaker is moved randomly while being accurately tracked by the robot.

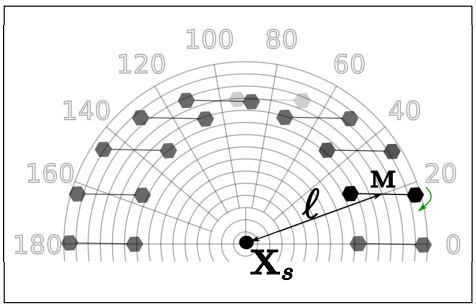
From the latter equation it can be seen that the tracking error due to the motion of the source can be compensated under the condition that

$$\mathbf{L_e}\widehat{\mathbf{L}_e^+}\widehat{\frac{\partial e}{\partial t}} = \frac{\partial e}{\partial t} \tag{4.52}$$

Several methods have been developed in control literature in order to estimate $\widehat{\frac{\partial e}{\partial t}}$. To name few of them, a Kalman filter is proposed in [CG93], while a predictive control scheme is used in [GdM02] and an integrator in [CH08].

4.2.6 Evaluation and limitations of the ILD-based task

In order to evaluate the performance of the ILD-based task, we tested the method through simulations that could guarantee repeatability and ground truth references to assess of the accuracy of the task. The simulation environment is based on *Room-simove* [BOV12], a room acoustics simulator, that allows to parametrize the room size, reverberation, and the location of the source and microphones. From this tool, we designed a room of 8×6 m^2 in which a sound source $\mathbf{X}_s$ is located. Similarly to the first experiment presented in this section, the simulated tasks consisted in facing the sound source from several poses of the robot as depicted in Figure 4.10. The task was repeated for several initial poses of the robot that are varied by 20° around the sound source for a given distance ℓ within the range $[0°, 180°]$. This decomposition yielded to 11 initial poses. It should be noted that the initial orientation of the robot is fixed, and only the initial positions are varied. We principally evaluated the orientation error of the final pose of the robot with respect to the desired task (facing the source). Such evaluation was repeated for several distance $\ell \in \{0.5, 1, 2, 3\}$ and several level of reverberation $\mathrm{RT}_{60} \in \{0, 50, 100, 200\}$ ms. The results are summarized in Figure 4.10, where the mean absolute error, in degree, is reported.



RT_{60} (s)	ℓ (m)			
	0.5	**1**	**2**	**3**
0	< 1	< 1	2.7	6.4
0.05	< 1	3.6	5.6	7.6
0.1	< 1	5.1	7.0	8.4
0.2	< 1	8.8	10.4	21.4

Figure 4.10: A simulated task that consists in orienting the robot in the direction of the sound source, from different poses in a 8×6 m^2 room. The final absolute mean error, in degree, is calculated for several reverberation rate (RT_{60}) and distances to the source.

The principal information that can be drawn from these simulations assesses that an ILD-based positioning task is accurate only in the close neighborhood of the sound source. This constraint can be explained by two distinct properties of ILDs that can be interpreted from the results. First the simulations confirm that increasing the distance between the microphones and the source degrades the accuracy of the task. This is confirmed by the simulations in anechoic conditions, from which it can be observed a decrease of accuracy as soon as the microphones get further to the source. This is caused by the limitations inherent to ILD definition $\rho = \frac{\ell_2^2}{\ell_1^2}$. From, the scene

configuration described in Figure 4.1, the relationship between ℓ_1 and ℓ_2 is given by

$$\ell_2^2 = \ell_1^2 + 2dx_s \tag{4.53}$$

while $\ell_1^2 = \ell^2 + \frac{d^2}{4} - dx_s$ and $\ell_2^2 = \ell^2 + \frac{d^2}{4} + dx_s$. When the robot is far from the microphones, ℓ (respectively each ℓ_i) becomes large in comparison with d, the inter-microphone distance. As a result, it comes out that $\ell_1^2 \to \ell_2^2$ and $\rho \to 1$. Thus the energy difference between the two microphones becomes too small and the dynamics of the ILD ρ is not significant anymore: large motions of the robot induced a small change in the ILD which explains the results in anechoic conditions. This result also confirms that ILD measurements in free-field conditions are more relevant for wide inter-microphones distance, which is unfortunately not compatible with robotic context. Paradoxically, head-mounted systems that are supposed to be little accurate, are less affected by this limitation (see Chapter 6), because of the additional sound attenuation of the head (at high frequencies).

Furthermore, as already stated in [BG05] or [CCW06], the results exposed in Figure 4.10 confirm that the ILD measurement is highly sensitive to reverberation. Systematic bias corrupts ρ when the microphones are far from the sound source, especially for positions close to walls. Indeed, with a reverberation modelling based on image source model [AB79], p virtual sources can be considered in the scene. Every virtual sound source j emits a sound wave so that each microphone $\mathbf{M_i}$ perceive an additional energy characterized by

$$E_p = \sum_{j=1}^{p} \frac{1}{\ell_{ij}^2} \int_{t=0}^{w} a^2(t - \frac{\Delta_{ij}}{c})\, \mathrm{d}t \tag{4.54}$$

where $\frac{\Delta_{ij}}{c}$ is the delay induced by the virtual source j. These virtual sound sources interfere with the signal received from the actual sound source. For the sake of simplicity, let us consider the upper bound of this additional energy, by assuming that (4.54) is a purely constructive sum, and that the source is very close to the reflective surfaces (i.e the virtual sources are not time-delayed w.r.t the actual source). In this "worst" case scenario, the measured ILD is modified as follow:

$$\rho = \frac{\frac{1}{\ell_1^2} + \sum_{j=1}^{p} \frac{1}{\ell_{1j}^2}}{\frac{1}{\ell_2^2} + \sum_{j=1}^{p} \frac{1}{\ell_{2j}^2}}. \tag{4.55}$$

This kind of erroneous measurement has a limited effect on the interaction matrix that is already approximated. Nonetheless, the accuracy of the convergence can be influenced by the error e that is equal to

$$e = \frac{\frac{1}{\ell_1^2} + e_{1p}}{\frac{1}{\ell_2^2} + e_{2p}} - \rho^*. \tag{4.56}$$

where $e_{1p} = \sum_{j=1}^{p} \frac{1}{\ell_{1j}^2}$ and $e_{2p} = \sum_{j=1}^{p} \frac{1}{\ell_{2j}^2}$. Therefore if e_{1p} and e_{2p} are significant enough w.r.t $\frac{1}{\ell_i^2}$, the error e of the control loop is not null. It can then be deduced

that a sufficient condition to obtain accurate ILD measurement (resp. accurate error e) is a high Direct-to-Reverberant Ratio (DRR).

Hence under realistic conditions, that is to say in presence of reverberation, and if the robot is far from the source, the performance of ILD-based positioning task might be drastically decreased. To overcome this limitation, we introduce the sound energy $E_{\mathbf{M}}$ as an additional feature corresponding to a distance cue. the dynamics of this cue can be integrated in the control scheme in order to regulate the distance to the sound source by setting a desired energy level. In this way, a positioning task could be performed accurately from an initial configuration relatively far from the source.

4.3 Integrating the absolute level of energy as distance cue

The distance of observation for accurate ILD is optimal in the near-field as demonstrated above. However the near-field assumption cannot be guaranteed in uncontrolled environments. In this ultimate section of the chapter, we introduce a distance cue that can bring the robot to the near-field zone. Namely, this cue is the absolute level of sound energy. The sound level is naturally a measure of distance. As stated in Chapter 2.2, it is complex to deduce a metric value of distance from this cue. Indeed the sound level depends on the intrinsic level of the emitting sources and is also dependent of the room acoustic that may add destructive or constructive interferences. The distance cues are not often used in robot audition because of these limitations although some studies showed that coarse approximations could be obtained by combining binaural cues to signal amplitude information [Rod10] or by estimating the DRR [LC10]. In contrast to approaches based on localization, AS allows us to exploit distance cues such as the absolute level of energy. In this paradigm, we are not attached to extract a given metric measurement related to the sound level but rather to use the dynamics of the sound energy, related to the sound decay, so as to position the robot at a given distance to the sound source.

4.3.1 Energy estimation

For a first step, let us analyze the consistency of sound energy with respect to the reverberation issue, through AS framework. The reverberation as denoted in the previous section drastically decreases the performance of the ILD-based positioning task from far-field. It is then necessary to prove that the absolute energy level is less affected by reverberation and thus can overcome this limitation. For more versatility and flexibility, we consider the measurement of the absolute energy in between the microphones, that is to say at the position $\mathbf{M}$. From (4.4), the sound energy $E_{\mathbf{M}}$ received by the point $\mathbf{M}$ is given by

$$E_{\mathbf{M}} = \frac{1}{\ell^2} \int_{t=0}^{w} a^2(t - \frac{\ell}{c})\, \mathrm{d}t. \tag{4.57}$$

In order to prove that reverberation has a minor effect on the sound energy, let us consider the upper limit reflective perturbations discussed above for the ILD. Assuming that $E^*_{\mathbf{M}}$ is experimentally measured beforehand, both $E_{\mathbf{M}}(t)$ and $E^*_{\mathbf{M}}$ include the error e_p caused by the wall reflections. Therefore the energy level error in the control loop is equal to

$$e_{\mathbf{M}} = \left(\frac{1}{\ell^2} + e_p\right)\int_{t=0}^{w} a^2(t - \frac{\ell}{c})\mathrm{d}t - \left(\frac{1}{\ell^{*2}} + e^*_p\right)\int_{t=0}^{w} a^2(t - \frac{\ell^*}{c})\mathrm{d}t \tag{4.58}$$

If $E^*_{\mathbf{M}}$ is measured so that the sensors are in the near-field (*i.e.*, high DRR), as it is targeted in order to overcome ILD limitations, one can deduce that $\frac{1}{\ell^{*2}} \gg e^*_p$. In the far-field it is expected that the level of sound energy is lower than in the near-field because of the sound decay phenomena detailed in Chapter 1.2. It can then be written that $\frac{1}{\ell^2} + e_p < \frac{1}{\ell^{*2}} - e^*_p$, with ℓ and e_p corresponding to a far-field measurement. While approaching the desired pose, we obtain:

$$e_{\mathbf{M}} \underset{s \to s^*}{=} \left(\frac{1}{\ell^2} - \frac{1}{\ell^{*2}}\right)\int_{t=0}^{w} a^2(t - \frac{\ell^*}{c})\,\mathrm{d}t. \tag{4.59}$$

Thus, the reverberation has a minor effect on the energy level cue, in the configuration in which the robot has to reach or stay in the near-field zone. From then on, since the energy level in $\mathbf{M}$ is not directly available, we just approximate $E_{\mathbf{M}}$ as the mean value of the energy received by each microphones:

$$E_{\mathbf{M}} \approx \frac{E_1 + E_2}{2}. \tag{4.60}$$

4.3.2 Energy level modelling

In a similar process as the one developed in Section 4.2 for modelling $\mathbf{L}_\rho$, the interaction matrix related to the absolute sound energy $\mathbf{L}_{E_{\mathbf{M}}}$ can be inferred. Let us start from (4.2), specifying the distance to the source as $\ell = \sqrt{x_s^2 + y_s^2}$. The time variation of this distance is defined by:

$$\dot{\ell} = \frac{x_s\dot{x}_s + y_s\dot{y}_s}{\ell}. \tag{4.61}$$

By injecting the kinematic equation (4.16) in (4.61), the interaction matrix related to ℓ is easily obtained as

$$\mathbf{L}_\ell = \begin{bmatrix} -\frac{x_s}{\ell} & -\frac{y_s}{\ell} & 0 \end{bmatrix}. \tag{4.62}$$

With the assumption that the source emits a constant signal during the time frame w, from (4.57) the time variation of $E_{\mathbf{M}}$ is

$$\dot{E_{\mathbf{M}}} = -2E_{\mathbf{M}}\frac{\dot{\ell}}{\ell}. \tag{4.63}$$

Therefore the interaction matrix $\mathbf{L}_{E_{\mathbf{M}}}$ related to the sound energy perceived in $\mathbf{M}$ is given by

$$\mathbf{L}_{E_{\mathbf{M}}} = -2\frac{E_{\mathbf{M}}}{\ell}\mathbf{L}_\ell = E_{\mathbf{M}}\begin{bmatrix} \frac{2x_s}{\ell^2} & \frac{2y_s}{\ell^2} & 0 \end{bmatrix}. \tag{4.64}$$

Once again, an approximation of the latter interaction matrix is necessary since the sound source location is unknown:

$$\widehat{\mathbf{L}_{E_\mathbf{M}}} = E_\mathbf{M} \begin{bmatrix} \frac{2\widehat{x}_s}{\widehat{\ell}^2} & \frac{2\widehat{y}_s}{\widehat{\ell}^2} & 0 \end{bmatrix}. \tag{4.65}$$

.

It can be noticed that the sufficient stability conditions of a task based on $\widehat{\mathbf{L}_{E_\mathbf{M}}}$ are the same as the ILD described in Section 4.2.4. Indeed

$$\mathbf{L}_{E_\mathbf{M}} \widehat{\mathbf{L}_{E_\mathbf{M}}^{+}} = \frac{\widehat{\ell}^2}{\ell^2} \frac{x_s \widehat{x}_s + y_s \widehat{y}_s}{\widehat{x}_s^2 + \widehat{y}_s^2}, \tag{4.66}$$

is positive as soon as $sign(\widehat{x}_s) = sign(x_s)$ and $sign(\widehat{y}_s) = sign(y_s)$. Hence $\widehat{x}_s$, $\widehat{y}_s$ and $\widehat{\ell}$ can be estimated with the same parameters used in $\widehat{\mathbf{L}_\rho}$.

4.3.3 Control scheme and task analysis

For the interaction matrix $\mathbf{L}_{E_\mathbf{M}}$ of the energy level, the subspace $\mathbf{S}^*$ related to the motions for which the energy stays constant is:

$$\mathbf{S}^* = \begin{bmatrix} 0 & -y_s \\ 0 & x_s \\ 1 & 0 \end{bmatrix}. \tag{4.67}$$

The first motion implied by $\mathbf{S}^*$ is a pure rotation. Indeed the distance ℓ between $\mathbf{M}$ and the source is invariant with respect to the orientation of the microphones. The second column illustrates all translations for which the distance ℓ to the sound source is unchanged. Namely, it refers to a circular motion around the sound source. Consequently by combining these two types of motion, $\mathbf{S}^*$ describes a circle of radius ℓ around the sound source with unconstrained orientation of the microphones, that is illustrated by Figure 4.11. This result emphasizes that the ILD and the energy level are complementary features. Indeed, in contrast with the energy level, the ILD constrains the orientation of the microphones while the distance is constrained by the energy level.

Then it can be defined an interaction matrix $\widehat{\mathbf{L}_{\rho E}}$ combining the ILD ρ to the energy level $E_\mathbf{M}$ by simply stacking $\widehat{\mathbf{L}_\rho}$ and $\widehat{\mathbf{L}_{E_\mathbf{M}}}$ as following:

$$\widehat{\mathbf{L}_{\rho E}} = \begin{bmatrix} \frac{2\widehat{x}_s(\rho-1)-d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{2\widehat{y}_s(\rho-1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{\widehat{y}_s d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} \\ \frac{2E_\mathbf{M}\widehat{x}_s}{\widehat{\ell}^2} & \frac{2E_\mathbf{M}\widehat{y}_s}{\widehat{\ell}^2} & 0 \end{bmatrix}. \tag{4.68}$$

From the latter interaction matrix, the control scheme related to the velocity of the microphones is defined as

$$\mathbf{u}_\mathbf{M} = -\lambda \widehat{\mathbf{L}_{\rho E}}^{+} \mathbf{e} \tag{4.69}$$

where the error vector is defined as $\mathbf{e} = [\rho - \rho^*, E_\mathbf{M} - E_\mathbf{M}^*]^\top$. For this control scheme integrating the energy level, the vector subspace $\mathbf{S}^*$ becomes a rank-one matrix given

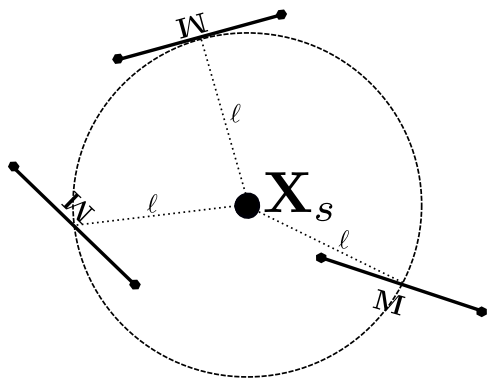


Figure 4.11: Admissible poses of the microphones considering the sound energy level as acoustic features

by

$$\mathbf{S}^* = \begin{bmatrix} y_s \\ -x_s \\ 1 \end{bmatrix}. \tag{4.70}$$

In this case, the motion described by this subspace refers to a circular motion around the sound source at a distance ℓ maintaining the same orientation of the microphones with respect to the source (see Figure 4.12). Consequently with a good selection of

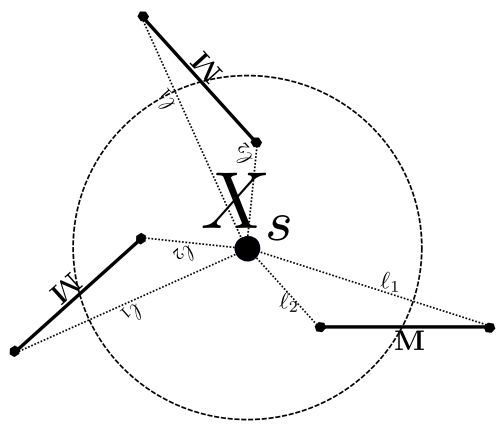


Figure 4.12: Admissible poses of the microphones considering the sound energy level and the ILD as acoustic features

the desired features value, it is possible to reach a pose around the sound source by only considering the measurement of the energy and the ILD. Such task can be referred to a coarse-to-fine approach. When the robot is far from the sound source, the ILD measurement just gives a rough approximation of the actual direction of the source. However as the robot is moving closer, the ILD measurement is refined, and the motion of the robot becomes more accurate. Hence, from rough and approximated features measurements, a virtuous cycle is created from the closed-loop control, that allows to achieve difficult tasks (see Section 4.3.4). Eventually, it should

be emphasized that until now, no additional signal processing, nor filtering have been employed for the control scheme. Additionally, the features measurements and the control scheme are independent from any additional tracking method, as long as the sound source of interest is predominant in the environment. The computation cost of this AS framework is drastically decreased, when compared to classic localization methods in adverse conditions. Actually, the most complex calculation of the framework consists in computing the pseudo-inverse of $\widehat{\mathbf{L}_{E_\mathbf{M}}} \in \mathbb{R}^{2\times 3}$.

4.3.4 Experimental validations

4.3.4.1 Preliminaries: control scheme and experimental setup

Once again, we consider the 2-DOF robot modelled in Section 4.2.5.1. The control input of this robot is given by

$$\dot{\mathbf{q}} = -\lambda \widehat{\mathbf{J}^{-1}_{\rho\mathbf{E}}} \mathbf{e} \tag{4.71}$$

with

$$\widehat{\mathbf{J}_{\rho\mathbf{E}}} = \begin{bmatrix} \frac{2\widehat{y}_s(\rho-1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{D_x(2\widehat{x}_s(\rho-1)-d(\rho+1))+\widehat{y}_s d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} \\ \frac{2E_\mathbf{M}\widehat{y}_s}{\widehat{\ell}^2} & \frac{2D_x E_\mathbf{M}\widehat{x}_s}{\widehat{\ell}^2} \end{bmatrix}. \tag{4.72}$$

Apart from the degenerate cases already mentioned previously ($\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s = 0$ and $\widehat{\ell}^2 = 0$), singularities of the control system could appear as soon as $\widehat{y}_s = D_x$ or $\widehat{y}_s = 0$. In these particular situations, the determinant of this Jacobian matrix given by

$$|\widehat{\mathbf{J}_{\rho\mathbf{E}}}| = \frac{2E_\mathbf{M}\widehat{y}_s d(D_x - \widehat{y}_s)(1+\rho)}{\widehat{\ell}^2(\widehat{\ell}^2 + \frac{d^2}{4} - d\widehat{x}_s)}. \tag{4.73}$$

would be equal to 0, and the interaction matrix would not be invertible. Hopefully, in practice the sound source position is approximated so that $\widehat{y}_s > D_x$.

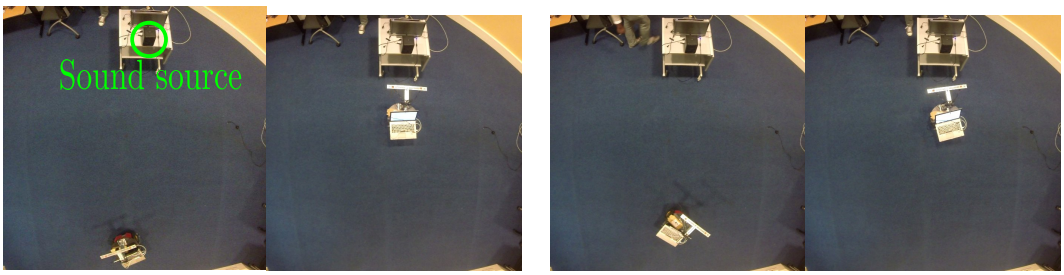
The experiments were conducted on the *Pioneer 3DX* with the same configuration given in Section 4.2.5.1 and illustrated in Figure 4.5. The global processing time of one iteration was also negligible (< 10 ms). The acoustics conditions are detailed in the sections devoted to each experiment. The parameters given in Figure 4.5 were used for all experiments. An adaptive gain $\lambda(x)$ in which x refers to the norm of the error $\mathbf{e}$ is used to smooth and speed up the robot motion.

d	0.31 m
D_x	0.3 m
$\widehat{y}_s$	1 m
$\widehat{x}_s$	sign$(\rho-1)$ $\times$1 m
λ	$0.45e^{(-1.5x)}$

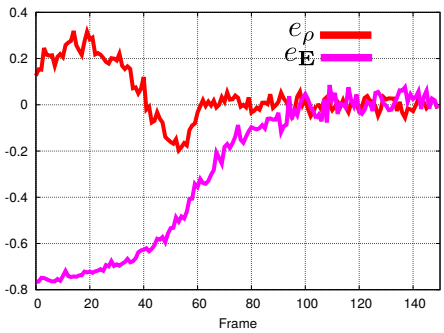
Table 4.2: Experimental settings

4.3.4.2 Typical positioning tasks

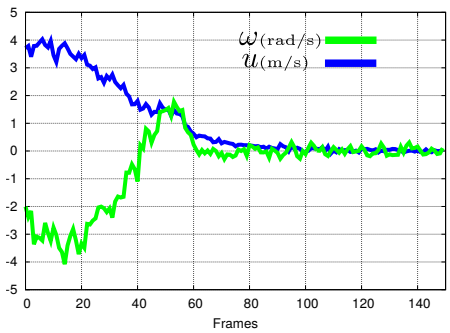
The first tests were conducted in a room characterized by a reverberation time $RT_{60} \approx 580$ ms. The sound source is a loudspeaker emitting a white Gaussian noise. The desired energy level is learned by placing the robot at 80 cm in front of the robot, while the desired ILD is set so that $\rho^* = 1$. The SNR at the desired pose is around 20 dB. The loudspeaker being directional, the admissible poses of the microphones were not in a circular configuration as stated in Figure 3.2, but they were shaped by the directivity properties of the speaker.



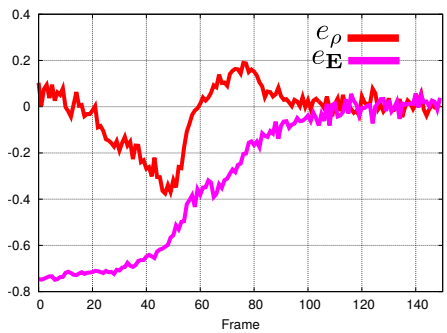
(a) Starting pose 1 (b) Final pose 1 (c) Starting pose 2 (d) Final pose 2



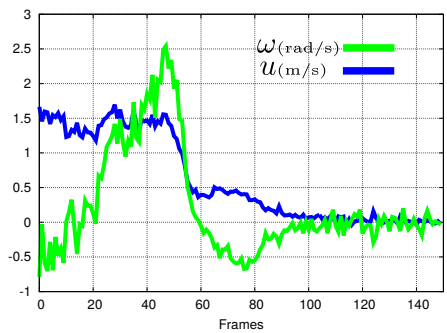
(e) Features error 1



(f) Velocities input 1



(g) Features Error 2

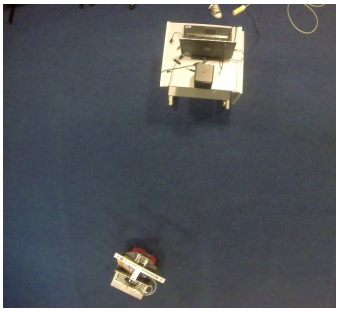


(h) Velocities input 2

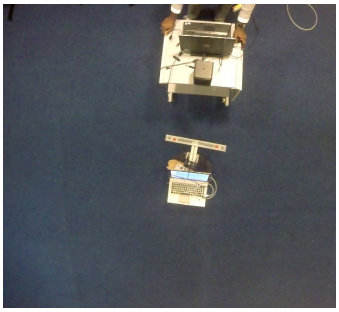
Figure 4.13: Typical positioning tasks from two different starting poses

First we showed the consistence of our approach with a static sound source, by starting the robot from different poses. As illustrated in Figure 4.13, the control

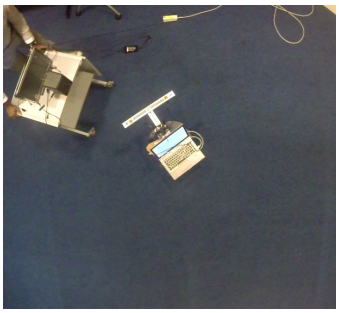
scheme allows to reach a pose satisfying the desired features from various initial poses. Actually, as long as the difference of energy between the microphones is perceptible at the initial pose, the control scheme is able to position the robot in a desired configuration. A fine tuning of λ or w can also improve the sensitivity of the control scheme to small differences in the energy level. By increasing the size of w, the difference of energy recorded by the pair of microphones is more significant. And λ tunes the time to convergence.



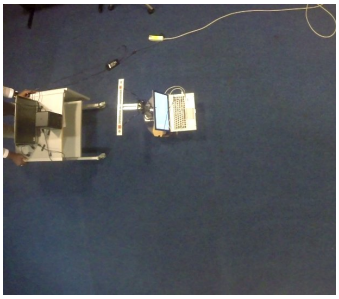
(a)



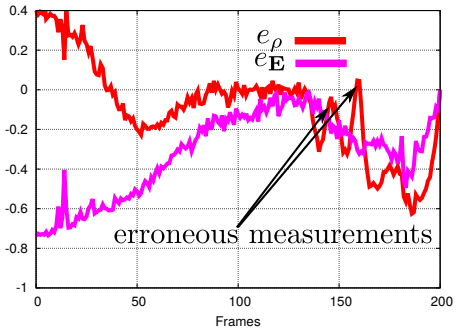
(b)



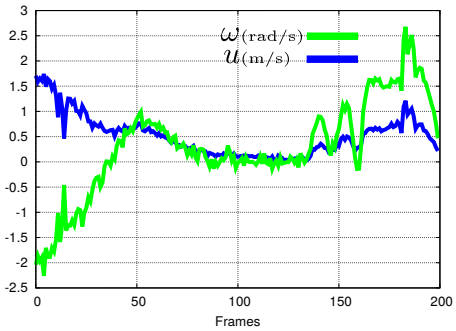
(c)



(d)



(e) Features error



(f) Velocities input

Figure 4.14: Following a moving sound source

In the second part of the experiment, we considered a moving sound source. As shown in Figure 4.14, the robot is able to follow accurately the sound source, with exactly the same control scheme and without any knowledge on the motion of the

sound source. As expected, when the robot is far from the sound source, it can be observed on Figure 4.14e that the ILD is not always accurate. Indeed we can notice some abrupt changes in the ILD error curve. But as soon as the robot gets closer to the sound source, the ILD value is correct, and the task can be correctly completed.

4.3.4.3 Robustness and flexibility in long range navigation

In a second experiment, we conducted a long range navigation test. Starting from the previous room, we moved the source through different environments of the laboratory with the robot pursuing the sound source in real time. Thus several acoustic conditions were encountered during the navigation as described in Figure 4.15.

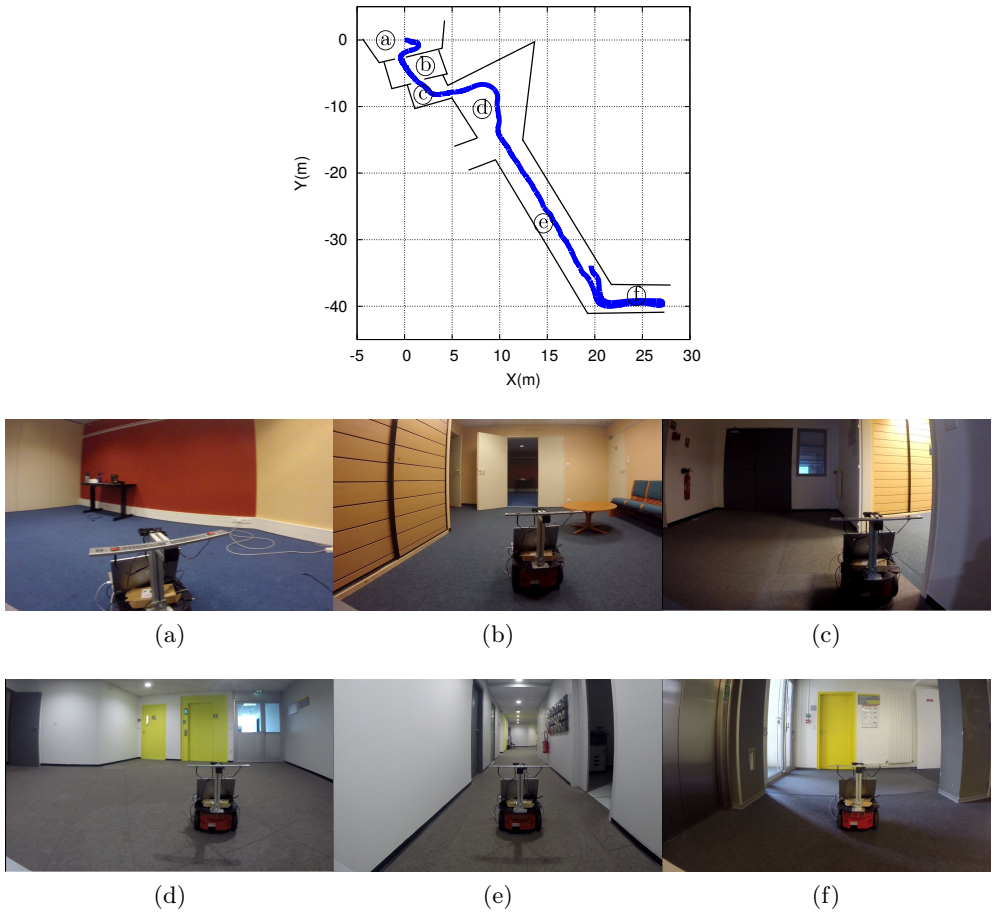


Figure 4.15: Odometry data from the navigation task in indoor environment.The acoustic conditions for each location are respectively:
ⓐ $RT_{60} \approx 580$ms SNR$\approx$ 20dB, ⓑ $RT_{60} \approx 620$ms SNR$\approx$ 20dB, ⓒ $RT_{60} \approx 680$ms SNR$\approx$ 16dB, ⓓ $RT_{60} \approx 880$ms SNR$\approx$ 13dB, ⓔ $RT_{60} \approx 700$ms SNR$\approx$ 18dB, ⓕ $RT_{60} \approx 620$ms SNR$\approx$ 17dB.

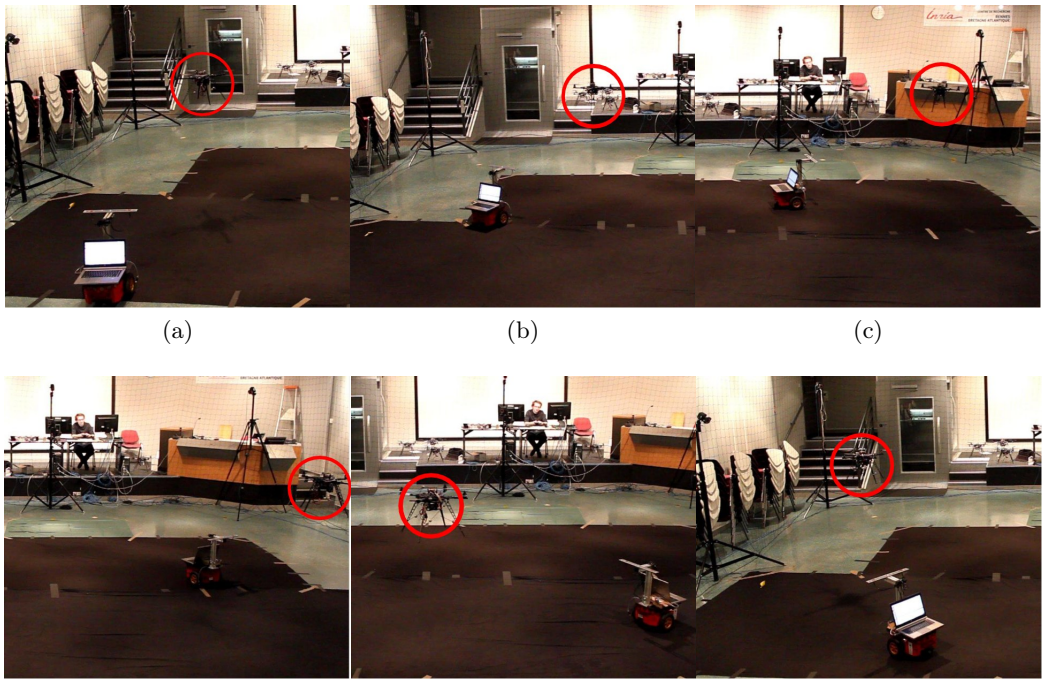
The SNR varied from 20 dB to 13 dB while the reverberation rate changed from 580 ms to 880 ms depending on the location. The ambient noise was mainly caused

by the ventilation systems and a server room in ⓓ. Nonetheless the robot never lost the track of the sound source despite dynamic and challenging conditions. Indeed the rooms and corridors crossed by the robot have different shapes, and are built with different materials and clutter. For instance the corridor ⓔ is narrow and produces strong first order echoes. Besides the echoes are not necessarily symmetric, because of the office doors that were open. In ⓕ different surrounding materials (glass door, metallic door of the elevator, walls...) were producing several types of echoes that could disturb the control scheme.

Consequently, from this experiment it can be emphasized that the proposed control scheme is robust and flexible enough to deal with indoor real world environment. Indeed, since our method does not rely on any tracking or filtering of the signal, there is no tuning nor parameter dependent on the acoustic environment. Thus our approach is robust to environment changes.

4.3.4.4 Cooperative application

The framework has also been tested in a cooperative task involving two robots. This time, instead of using the loudspeaker to generate the sound, we used the propellers of an unmanned aerial vehicle (UAV) as the sound source. Indeed most UAVs are known to be noisy. However, in a classic SSL scheme, this sound is considered more as a disturbance than a feature to exploit [BSFL14][FON+13], while our approach takes advantage of this noise. In this experiment an UAV (mikrokopter MK-Quadro), remotely controlled, led the unicycle ground robot just by the sound naturally emitted from the propellers. Nonetheless it should be noted that the sound emitted by the UAV was not stationary nor omnidirectional. Indeed the UAV was oscillating during the flight and the sound was produced by the four propellers of the UAV. Nevertheless, despite these unfavourable conditions, the ground robot was able to follow the UAV (see Figure 4.16), even if the motion of the robot was less smooth than in the previous experiments. This basic experimental scenario confirmed that our approach is relevant and suitable for cooperative tasks involving several robots. Furthermore, the control scheme based on only two microphones and the low computation cost of the framework make us believe that this kind of control scheme can be embedded on different type of robots. More evolved and complex tasks could then be achieved involving cooperation between aerial and ground robots or formation control of swarm robots.



(a) (b) (c)

(d) (e) (f)

Figure 4.16: An example of cooperative application where an aerial robot (circled in red) is leading a ground robot with the sound emitted by its propellers

4.3.5 Numerical evaluation

Eventually, the control scheme was tested through simulations. The simulations environments and conditions are strictly identical to the evaluation performed in Section 4.2.6. This time, the task consists in approaching, at a given distance, the sound source in order to face it. Hence ρ^* is set to 1, while $E^*_{\mathbf{M}}$ is measured experimentally in anechoic conditions as the energy measurement when $\rho = 1$ and $\ell = 0.5$ m. It should be emphasized that $E^*_{\mathbf{M}}$ has been measured only once, and the same value of $E^*_{\mathbf{M}}$ is used for all acoustic conditions. The results are summarized in Table 4.3, where the mean absolute error of orientation, in degree, and the mean absolute error of the range positioning, in cm, are reported. These results show impressive improvements compared to the task of only facing the sound source. For all starting poses, even in the far-field and with reverberation, the robot is always able to face to the sound source accurately as the errors reported are all lesser than 1°. It can be also noted that as expected and demonstrated in Section 4.3.1, the energy level as range cue is little affected by reverberation, with negligible errors varying from 1 to 4 cm under moderate reverberation. Actually, our control scheme for range positioning outperforms the solution based on localization proposed in [Rod10] or [LC10] that can estimate, at best, the distance with an approximate error of 36 cm for the first one and 1 m (with $\mathrm{RT}_{60} = 0.2$ s) in the second work. Furthermore, interestingly, the energy level is consistent over changing conditions.

$\mathrm{RT}_{\mathbf{60}}$ (s)	$\boldsymbol{\ell}$ (m)							
	0.5		**1**		**2**		**3**	
0	< 1	(1.47)	< 1	(1.47)	< 1	(1.48)	< 1	(1.47)
0.05	< 1	(1.77)	< 1	(1.77)	< 1	(1.78)	< 1	(1.8)
0.1	< 1	(1.94)	< 1	(2.05)	< 1	(1.95)	< 1	(1.96)
0.2	< 1	(3.90)	< 1	(3.90)	< 1	(3.96)	< 1	(3.98)

Table 4.3: A simulated task that consists in approaching and facing of the sound source ($\rho^* = 1$, $E^*_{\mathbf{M}} \equiv \ell^* = 0.5$ m), from different poses in a 8×6 m^2 room. The final absolute mean error, in degree, and the final range error (bracketed values), in cm, are calculated for several reverberation rate (RT_{60}) and distances to the source.

Despite an initial measurement $E^*_{\mathbf{M}}$ performed under anechoic conditions, the range error remains lower than 5 cm for a reverberation level $\mathrm{RT}_{60} = 0.2$ s. Of course, more accurate results could be obtained if $E^*_{\mathbf{M}}$ is measured in the same acoustic conditions as the task environment. These excellent properties substantially explain the good results obtained in experiments, in which our approach is robust and flexible to changing and real world environments.

4.4 Conclusion

This chapter globally emphasizes several benefits of the AS paradigm. First, the two solutions presented are based on features (ILD, energy as distance cue), that are challenging to exploit under binaural localization paradigm. For instance most of the works exploiting ILDs conjointly estimate ITDs, in order to map them with an azimuth direction from a learning process [MvdPK11]. However, these cues can be fully exploited with a control scheme based on their dynamics.

To do so, we evaluated geometrically the information gathered by ILD measurements. From ILD measurements, the sound source location is expressed through a circle depending on the ratio of energy. This circle is subject to the *front-back ambiguity*, since the ILD circle is centred somewhere on the extension of the microphones interaural axis. As a second part, we focused on the controller of the sensor-based scheme, more exactly the relationship between the dynamics of the ILD with respect to the potential motion of the robot. As a result, an approximated interaction matrix could be extracted from an ILD cue. From the design of a sensor-based control scheme, the task analysis confirmed that ILD information could be exploited to orient the robot with respect to the position of the sound source. Eventually the stability analysis showed that the source location is not required for this positioning task. Actually, there exists infinite number of solutions that guarantees the convergence of the controller to complete such tasks.

The second interesting benefit of AS comes from its ability to deal with real world

conditions thanks to the flexibility and the robustness of the approach. The framework described above has been validated experimentally in real conditions despite an initial modelling based on ideal conditions. In our experiments conducted on a mobile robot, we showed that from our approach a robot is able to face accurately a sound source, to track a moving sound source, and more impressively to cope with *front-back ambiguity*, without any information about the location of the source.

In the second part of this chapter, we integrated a distance cue to our model to overcome the limitation inherent to ILD: accuracy decreases for a distant source. The distance cue is the energy level measured by the microphones. The ILD allows orienting the robot towards the sound source while the signal energy controls the distance between the robot and the sound source. By combining these two features the robot is able to follow a moving sound source, despite erroneous ILD measurements, as confirmed in the experiments.

Furthermore, this framework is directly performed on raw measurements without tracking, filtering or signal enhancement. As a result, the computational cost of the control scheme is small, less than 10 ms in our experiments. The different experiments described in this chapter emphasized the robustness of the control scheme towards reverberation, variability of the environment. The different use cases showed the applicability of the method in various topic that could take advantage of the auditory sense:

- A mobile robot navigated by pursuing a sound source through indoor environment while facing different level of reverberation and noise. This kind of tasks has potential applicability to a wide a range of field such as security patrolling or delivery services.

- An aerial robot has been able to guide a ground robot only with the sound of its propellers. In this case, applications in the field of multi-robot and cooperative tasks, or search-and-rescue mission can be targeted.

Globally the ILD-based control framework presented in this chapter could help fulfilling the requirements for robots endowed with hearing sense that are: embeddability, real time processing, flexibility to broad environments, robustness to noise and reverberation. But on the other hand, one constraint of this framework is the use of a continuous and slowly varying sound signal. Hence, the control framework is not particularly adapted to non-stationary, intermittent sources such as speech. Furthermore, until now we assumed a single source in the acoustic scene. For acoustic scenes considering speech or multiple sources, ITDs are more relevant cues. These cues are the subject of the next chapter that introduces an ITD-based positioning task framework.

Chapter 5

ITD-based aural servo

In the continuity of the previous chapter centered on ILDs, we introduce in this chapter the modelling of a sensor-based control framework using ITD measurements. ITDs are certainly the most studied auditory cues for localization in machine hearing. With a wide variety of techniques [PDA12, NMOK03] and microphones topology [VMRL03, KND06], the estimation of this cue concentrated a lot of effort from the robot audition community. However the results remain contrasted when facing dynamic realistic environments. Generally each method for estimating ITDs is tuned for a particular environment and context. By contrast, this chapter demonstrates that in AS paradigm, ITD cues can still be exploited despite changing environment or rough approximations in the acoustic scene modelling. The control framework is based on a relatively common ITD estimation algorithm: GCC-PHAT.

In a binaural configuration characterized by a pair of microphones in a free-field area, a geometrical analysis is performed in Section 5.1 in order to establish the link between an ITD measurement and the sound source location.

In Section 5.2, from this analysis, two interaction matrices are extracted from the dynamics of the ITD with respect to the motion of the microphones. These matrices are respectively extracted from the geometrical model of ITDs and the far-field assumption commonly adopted in the sound source localization. The control scheme and the positioning task are then analyzed from these interaction matrices. Thereafter, the stability properties, in Lyapunov sense, are studied to demonstrate the feasibility of such an approach from the control point of view. Eventually experimental results and numerical evaluations support the robustness and flexibility properties of the proposed AS framework.

In the last segment of this chapter, we focus in Section 5.3 on the context of multi-sources configuration. Control schemes considering two, three or more sound source are given. In this configuration, the positioning tasks refer to bearing-only homing systems. Experimental results and numerical evaluations are provided as well.

5.1 ITD modelling

5.1.1 Scene configuration

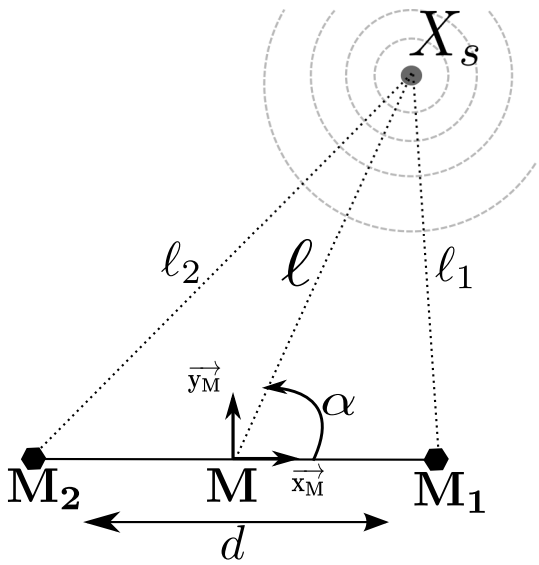


Figure 5.1: Geometric configuration of the considered system, that includes a source $\mathbf{X_s}$ emitting a spherical and uniform sound wave, and a pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$.

In a first phase, let us start from the configuration used for the ILD. We then recall the scene configuration that assumes a motorized pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ in an area free of obstacle and separated by a distance d illustrated in Figure 5.1. An omni-directional sound source $\mathbf{X_s}$ is continuously emitting. In the following, $\mathbf{X_s}$ is shaped as a point that generates a sound wave uniformly in all directions. A frame $\mathcal{F}_m(\overrightarrow{x_M}, \overrightarrow{y_M})$ is attached to the midpoint of the microphones $\mathbf{M}$. The Cartesian coordinates of each microphone are respectively $\mathbf{M_1}(\frac{d}{2}, 0)$ and $\mathbf{M_2}(-\frac{d}{2}, 0)$. The sound source $\mathbf{X_s}(x_s, y_s)$ is located at a distance ℓ_i from each microphone $\mathbf{M_i}$. These distances ℓ_i are respectively given by

$$\begin{cases} \ell_1 = \sqrt{(x_s - d/2)^2 + y_s^2} \\ \ell_2 = \sqrt{(x_s + d/2)^2 + y_s^2} \end{cases} \tag{5.1}$$

Moreover, let ℓ be the distance between $\mathbf{M}$ and $\mathbf{X_s}$, while α is the incident angle of the sound source with respect to the microphones axis. α is also known as the direction of arrival (DOA) of the source. The sound source position is thus characterized by the following relationships:

$$\begin{cases} x_s = \ell \cos \alpha \\ y_s = \ell \sin \alpha \end{cases} \text{and} \begin{cases} \alpha = \text{atan2}\,(y_s, x_s) \\ \ell = \sqrt{x_s^2 + y_s^2} \end{cases} \tag{5.2}$$

This configuration is then endowed with 3 DOF in the horizontal plane, that are the translations along $\overrightarrow{x_M}$ and $\overrightarrow{y_M}$ axis, and the rotation around $\overrightarrow{z_M}$.

5.1.2 Geometrical properties of the ITD

The first part of this chapter is dedicated to the study of the ITD, that corresponds to the time difference τ in sound arrival between two microphones. Let us focus first on the properties of this cue in the geometric configuration described above. From the Figure 5.1, the path of the signal emitted by $\mathbf{X_s}$ reaches each microphones $\mathbf{M_i}$ at a time t_i given by

$$t_i = \frac{\ell_i}{c}, \tag{5.3}$$

in which c is the sound celerity. As a result the ITD τ between the pair of microphones is given by

$$\tau = t_2 - t_1 = \frac{\ell_2}{c} - \frac{\ell_1}{c} \tag{5.4}$$

From the measurement of τ and the knowledge of the latter relationship, one can try to deduce the sound location. For this purpose, by injecting (5.1) in (5.4) we obtain:

$$\tau = \frac{\sqrt{(x_s + \frac{d}{2})^2 + y_s^2} - \sqrt{(x_s - \frac{d}{2})^2 + y_s^2}}{c}. \tag{5.5}$$

The latter equation develops as

$$\frac{\sqrt{(x_s + \frac{d}{2})^2 + y_s^2}}{c\tau} = \frac{c\tau + \sqrt{(x_s - \frac{d}{2})^2 + y_s^2}}{c\tau}, \tag{5.6}$$

that is equivalent to

$$\frac{(x_s + \frac{d}{2})^2 - (x_s - \frac{d}{2})^2}{(c\tau)^2} - 2\sqrt{\frac{(x_s - \frac{d}{2})^2 + y_s^2}{(c\tau)^2}} = 1. \tag{5.7}$$

After simplification we obtain

$$4\left(\frac{x_s^2(d^2 - (c\tau)^2)}{(c\tau)^2} - y_s^2\right) = d^2 - (c\tau)^2. \tag{5.8}$$

Finally the equation is reduced to

$$\frac{x_s^2}{\frac{(c\tau)^2}{4}} - \frac{y_s^2}{\frac{d^2-(c\tau)^2}{4}} = 1. \tag{5.9}$$

This equation corresponds to an hyperbola $\mathcal{H}$. Hence there are infinite solutions to characterize the potential location as depicted in Figure 5.2. It should also be mentioned that the same reasoning can be developed for 3D scenes (see [APH14]), in which $\mathcal{H}$ becomes an hyperboloid.

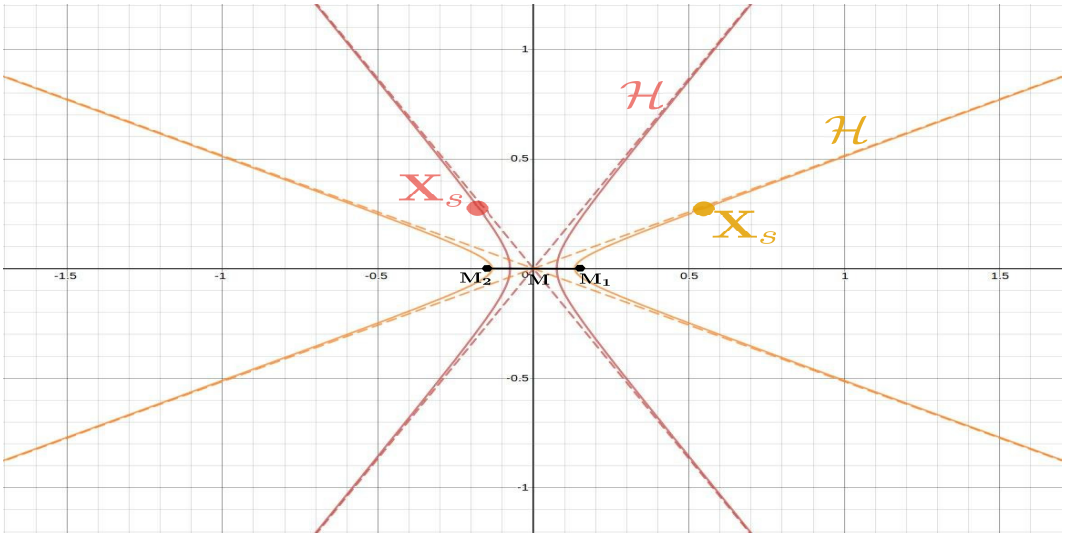


Figure 5.2: For a given sound source $\mathbf{X}_s$, the ITD geometrically refers to an hyperbola $\mathcal{H}$, for which front/back and left/right ambiguity should be solved. The asymptotes values (dotted line) of the hyperbola can serve as an approximation (far-field assumption) of the sound direction.

Several properties result from (5.9) in the planar case. First the hyperbola exists only if

$$-\frac{d}{c} < \tau < \frac{d}{c} \text{ and } \tau \neq 0. \tag{5.10}$$

As a result when $\tau = 0$, the hyperbola $\mathcal{H}$ degenerates to a straight line that corresponds to the bisection of the microphone pair (*i.e.* $x_s = 0$). This geometrical representation states that the sound source is "somewhere" on the hyperbola $\mathcal{H}$, or on the bisection of the microphones when the ITD $\tau = 0$. The position of the sound source can be refined by resolving the ambiguities related to $\mathcal{H}$. First, there is an ambiguity about determining on which branch of the hyperbola the sound source is located, for a given ITD. Fortunately, this ambiguity can be solved from the sign of the ITD τ. Indeed from (5.1), (5.7) can be rewritten

$$\frac{2x_s d}{(c\tau)^2} - \frac{2\ell_1}{c\tau} = 1. \tag{5.11}$$

With some rearrangements, (5.11) becomes

$$x_s = \frac{c\tau(c\tau + 2\ell_1)}{2d}. \tag{5.12}$$

Knowing from (5.4) that $c\tau = \ell_2 - \ell_1$, the sound source lateral location with respect to the microphones frame is determined by

$$x_s = \frac{c\tau(\ell_1 + \ell_2)}{2d}. \tag{5.13}$$

In plain, if $\tau < 0$ (*i.e.* the source is closer to $\mathbf{M_2}$ than $\mathbf{M_1}$), the source is located on the left branch, else the source is on the right branch. As for the front or back position of the sound source with respect with the microphones axis, it can be disambiguated from the sign of the hyperbola $\mathcal{H}$ asymptotes given by

$$y = \pm \frac{\sqrt{d^2 - (c\tau)^2}}{c\tau} x \tag{5.14}$$

However, the sign of asymptote is independent from the sign of the ITD and hence cannot be determined. This indeterminacy draws the already discussed concept of *front-back ambiguity*. By extension, in 3D scenes, this ambiguity related to the hyperboloid leads to the so-called cone of confusion inherent to mammals sound localization as discussed in Chapter 2.1.2.

Eventually, one can also emphasize that for distant sound sources, that is to say, for a source located at a pose where the hyperbola $\mathcal{H}$ has reached its asymptote value (see Figure 5.2), (5.14) can be written as

$$\tan \alpha = \pm \frac{\sqrt{d^2 - (c\tau)^2}}{c\tau}, \tag{5.15}$$

in which α is the DOA (see Figure 5.1). From trigonometry properties, one can develop (5.15) as

$$\frac{1 - \cos^2 \alpha}{\cos^2 \alpha} = \frac{d^2 - (c\tau)^2}{(c\tau)^2}, \tag{5.16}$$

that can be simplified as

$$(c\tau)^2 = d^2 \cos \alpha^2. \tag{5.17}$$

Finally one obtains

$$\tau = A \cos \alpha \tag{5.18}$$

where $A = d/c$. Equation (5.18) corresponds to the definition of the ITD with respect to the DOA under the far-field assumption. Thus it is possible to geometrically differentiate the near-field from the far-field by comparing how close the hyperbola is from its asymptotes, by using a distance criterion. Additionally, this geometrical modelling stresses the limitation of the far-field assumption that is commonly assumed in sound source localization (*e.g.* planar wavefront) [VMRL03, NOK01]. Indeed, the accuracy of the DOA obtained from the ITD decreases as the source gets closer to the microphones. Hence, rigorously our modelling should be based on (5.4), as it will be seen in Section 5.2.2, in order to control the robot accurately, independently of the distance to the sound source. Nonetheless, in the following, we also derive a model based on the ITD under the far-field approximation in (5.18). Although the far-field assumption cannot always be guaranteed in the scene, this kind of model may highlight one of the key points of the sensor-based control scheme: these two models may be equivalent from a control point of view. This property will be discussed exhaustively in Section 5.2.6.1.

5.2 An ITD-based interaction with a single source

5.2.1 ITD interaction matrix with far-field assumption

For now on, this section is concerned about linking the motions of the robot to the ITD dynamics. Indeed building the interaction matrix is the foundation of the sensor-based framework. In the configuration discussed above, one can start the model from the DOA α that is connected to the ITD τ through (5.18). The angle α can be expressed with respect to $\mathbf{X_s}(x_s, y_s)$ as :

$$\alpha = \operatorname{atan2}(y_s, x_s). \tag{5.19}$$

Thus the time derivative of α can be expressed as

$$\dot{\alpha} = \frac{\dot{y}_s x_s - \dot{x}_s y_s}{\ell^2}. \tag{5.20}$$

Equation (5.20) can also be expressed as a matrix relationship:

$$\dot{\alpha} = \mathbf{L}_\alpha \mathbf{v_M} \tag{5.21}$$

where $\mathbf{L}_\alpha$ is the interaction matrix related to α and $\mathbf{v_M} = (v_x, v_y, v_z, \omega_x, \omega_y, \omega_z)$ the spatial velocity of $\mathcal{F}_m$. Once again we define $\mathbf{u_M} = (v_x, v_y, \omega_z)$ that corresponds to the controlled DOF of the robot as described in Figure 5.1. From the kinematic equation

$$\dot{\mathbf{X}}_\mathbf{s} = -\boldsymbol{v_s} - \boldsymbol{\omega_s} \times \mathbf{X_s} \Leftrightarrow \begin{cases} \dot{x}_s = -v_x - \omega_y z_s + \omega_z y_s \\ \dot{y}_s = -v_y - \omega_z x_s + \omega_x z_s \\ \dot{z}_s = -v_z - \omega_x y_s + \omega_y x_s \end{cases} \tag{5.22}$$

we obtain

$$\dot{\alpha} = \frac{-v_y x_s + v_x y_s - \omega_z (y_s^2 + x_s^2)}{x_s^2 + y_s^2}. \tag{5.23}$$

In this equation, we can notice that $v_z = 0$, $\omega_x = 0$, $\omega_y = 0$, which limits the motion induced by the interaction matrix in the azimuth plane. Eventually referring to (5.21), the interaction matrix related to α is identified as:

$$\mathbf{L}_\alpha = \begin{bmatrix} \frac{y_s}{\ell^2} & -\frac{x_s}{\ell^2} & -1 \end{bmatrix} \tag{5.24}$$

in order to be consistent with $\mathbf{u_M}$ (only the non-zero terms are considered). Thereafter, since the sound source position $\mathbf{X_s}(x_s, y_s)$ is unknown, one can reshape the interaction matrix by using (5.2) so that $\mathbf{L}_\alpha$ becomes

$$\mathbf{L}_\alpha = \begin{bmatrix} \frac{\sin\alpha}{\ell} & -\frac{\cos\alpha}{\ell} & -1 \end{bmatrix}. \tag{5.25}$$

This result can subsequently be exploited to obtain the interaction matrix related to τ. Under the far-field assumption, the time derivative of (5.18) is:

$$\dot{\tau} = -A \sin\alpha \, \dot{\alpha}. \tag{5.26}$$

The interaction matrix $\mathbf{L}_{\tau_f}$ is then given by

$$\mathbf{L}_{\tau_f} = -A\sin\alpha\mathbf{L}_\alpha = \begin{bmatrix} -\frac{A\sin^2\alpha}{\ell} & \frac{A\sin\alpha\cos\alpha}{\ell} & A\sin\alpha \end{bmatrix}. \tag{5.27}$$

From (5.18) and the application of the trigonometric properties, the interaction matrix $\mathbf{L}_{\tau_f}$ is eventually reduced to

$$\mathbf{L}_{\tau_f} = \begin{bmatrix} -\frac{\nu^2}{A\ell} & \frac{\tau\nu}{A\ell} & \nu \end{bmatrix} \tag{5.28}$$

where $\nu = \sqrt{A^2 - \tau^2}$. Similarly to binaural localization, the distance ℓ that appears in $\mathbf{L}_\alpha$ and $\mathbf{L}_{\tau_f}$ is unknown. The actual interaction matrices related to τ or α are approximated as follows:

$$\widehat{\mathbf{L}_\alpha} = \begin{bmatrix} \frac{\sin\alpha}{\widehat{\ell}} & -\frac{\cos\alpha}{\widehat{\ell}} & -1 \end{bmatrix} \text{ and } \widehat{\mathbf{L}_{\tau_f}} = \begin{bmatrix} -\frac{\nu^2}{A\widehat{\ell}} & \frac{\tau\nu}{A\widehat{\ell}} & \nu \end{bmatrix}. \tag{5.29}$$

This kind of approximation is comparable to image-based visual servoing where the depth of a point, which is unknown, appears in the interaction matrix of an image point. An overview on these two matrices highlight the non-admissible values for the approximations. Indeed, for both matrices $\widehat{\ell} \neq 0$ is a necessary condition to avoid elements of infinite values in the interaction matrix. Infinite values would nullify the control input. In the same way, the relation $|\tau| < A$ related to the existence of the hyperbola $\mathcal{H}$ appears in the interaction matrix of τ as a condition of non-singularity ($i.e$ $\mathbf{u_M} = \infty$).

5.2.2 ITD interaction matrix without assumption

An interaction matrix $\mathbf{L}_{\tau_r}$ can also be obtained from the exact modelling of τ given by (5.4). That is to say $\mathbf{L}_{\tau_r}$ is computed without any assumption about the sound source being in the far-field or near-field. In this case the time derivative of τ is

$$\dot{\tau} = \frac{1}{c}(\dot{\ell}_2 - \dot{\ell}_1) \tag{5.30}$$

By injecting (5.1) in (5.30), the derivative of τ develops as

$$\dot{\tau} = \frac{1}{c}\left(\frac{2\dot{x}_s x_s + 2\dot{y}_s y_s + d\dot{x}_s}{2\ell_2} - \frac{2\dot{x}_s x_s + 2\dot{y}_s y_s - d\dot{x}_s}{2\ell_1}\right). \tag{5.31}$$

By using the kinematic equation given in (5.22), the latter equation becomes

$$\dot{\tau} = v_x\frac{x_s\tau - \frac{A}{2}(\ell_1+\ell_2)}{\ell_1\ell_2} + v_y\frac{y_s\tau}{\ell_1\ell_2} + \omega_z\frac{\frac{A}{2}(\ell_1+\ell_2)y_s}{\ell_1\ell_2}. \tag{5.32}$$

Finally $\mathbf{L}_{\tau_r}$ can be extracted from (5.32):

$$\mathbf{L}_{\tau_r} = \begin{bmatrix} \frac{x_s\tau - \frac{A}{2}(\ell_1+\ell_2)}{\ell_1\ell_2} & \frac{y_s\tau}{\ell_1\ell_2} & \frac{\frac{A}{2}(\ell_1+\ell_2)y_s}{\ell_1\ell_2} \end{bmatrix}. \tag{5.33}$$

Similarly to (5.29), unknown parameters depending on ℓ (x_s, y_s and ℓ_i are all function of ℓ and α) and thus related to the sound source location, appears in $\mathbf{L}_{\tau_r}$. The approximated feature interaction matrix is then

$$\widehat{\mathbf{L}_{\tau_r}} = \begin{bmatrix} \frac{\widehat{x}_s\tau - \frac{A}{2}(\widehat{\ell}_1+\widehat{\ell}_2)}{\widehat{\ell}_1\widehat{\ell}_2} & \frac{\widehat{y}_s\tau}{\widehat{\ell}_1\widehat{\ell}_2} & \frac{\frac{A}{2}(\widehat{\ell}_1+\widehat{\ell}_2)\widehat{y}_s}{\widehat{\ell}_1\widehat{\ell}_2} \end{bmatrix}. \tag{5.34}$$

For this type of interaction matrix, any approximation should respect the condition $\widehat{\ell}_i \neq 0$ in order to avoid elements of infinite values.

It should also be noticed that this interaction matrix (5.33) can easily be linked to the interaction matrix under the far-field assumption (5.28). Under the far-field assumption, the following hypothesis $\ell_1 \approx \ell_2 \approx \ell$ holds. Hence (5.33) can be rewritten as

$$\mathbf{L}_{\tau_r} \simeq \begin{bmatrix} \frac{x_s\tau - A\ell}{\ell^2} & \frac{y_s\tau}{\ell^2} & \frac{y_s A\ell}{\ell^2} \end{bmatrix}. \tag{5.35}$$

Afterwards, by replacing τ, x_s and y_s in (5.35) using (5.18) and (5.2), the interaction matrix becomes

$$\mathbf{L}_{\tau_r} \simeq \begin{bmatrix} -\frac{A\sin^2\alpha}{\ell} & \frac{A\sin\alpha\cos\alpha}{\ell} & A\sin\alpha \end{bmatrix} = \mathbf{L}_{\tau_f} \tag{5.36}$$

that corresponds to (5.27), the interaction matrix with a far-field assumption.

5.2.3 Control scheme

These interaction matrices can then be used to design a control scheme that consists in positioning the robot so that a given condition characterized by τ^* is satisfied. A positioning task is performed by considering a single ITD measurement $\tau(t)$ extracted from the sound signal and by minimizing the error $\|e(t)\|$. Hence this task is characterized by an error

$$e(t) = \tau(t) - \tau^* \tag{5.37}$$

where τ^* denotes the measurements for the desired ITD value. Similarly to the ILD case, we design a control scheme governed by the approximation of the interaction matrix $\widehat{\mathbf{L}_{\tau_j}}$, $j \in \{r, f\}$ in (5.29) or (5.34), in which the velocity of the microphones is computed from (5.21) as (see Chapter 3)

$$\mathbf{u_M} = -\lambda\widehat{\mathbf{L}^+_{\tau_j}}e. \tag{5.38}$$

We recall that $\lambda > 0$ is the gain that tunes the time to convergence. In the latter equation $\widehat{\mathbf{L}^+_{\tau_j}} \in \mathbb{R}^{3\times 1}$ is the Moore-Penrose pseudo-inverse of $\widehat{\mathbf{L}_{\tau_j}}$ (see Chapter 3.2.1). Thus, when considering the far-field assumption, the control scheme can be explicitly written as

$$\mathbf{u_M} = -\lambda\frac{\nu^2(1+\widehat{\ell}^2)}{\widehat{\ell}^2}\begin{bmatrix} -\frac{\nu^2}{A\widehat{\ell}} \\ \frac{\tau\nu}{A\widehat{\ell}} \\ \nu \end{bmatrix}(\tau - \tau^*). \tag{5.39}$$

Otherwise, when considering the more general interaction matrix given in (5.34), the control scheme becomes

$$\mathbf{u_M} = \frac{-\lambda\widehat{\ell}_1\widehat{\ell}_2}{(\widehat{y}_s\tau)^2 + (\frac{A}{2}(\widehat{\ell}_1+\widehat{\ell}_2)\widehat{y}_s)^2 + (\widehat{x}_s\tau - \frac{A}{2}(\widehat{\ell}_1+\widehat{\ell}_2))^2} \begin{bmatrix} \widehat{x}_s\tau - \frac{A}{2}(\widehat{\ell}_1+\widehat{\ell}_2) \\ \widehat{y}_s\tau \\ \frac{A}{2}(\widehat{\ell}_1+\widehat{\ell}_2)\widehat{y}_s \end{bmatrix} (\tau - \tau^*). \tag{5.40}$$

5.2.4 Task analysis

We analyze, here, the ITD-based positioning from (5.25), that offers more flexibility and easier to interpret. From this interaction matrix it is possible to achieve positioning tasks so that α reaches a particular desired value α^*. By using only one feature from a single sound source, it can be intuitively expected that several sensor poses exist so that $\alpha = \alpha^*$, that is a direct consequence of ITDs properties detailed previously. Actually only one DOF can be controlled by using one sound source. However, it is possible to extend the approach to control more DOF by simply increasing the number of sound sources as it will be performed for Section 5.3. Here we are attached to describe this task geometrically from virtual linkages approach given in Chapter 3.

This analysis is based on the vector subspace $\mathbf{S}^*$, where each column characterizes a motion for which α stays constant. When considering only one sound source, $\mathbf{L}_\alpha$ given by (5.25) is a rank-one matrix. Consequently $\mathbf{S}^* \in \mathbb{R}^{3\times 2}$ implies a virtual link of class 2:

$$\mathbf{S}^* = \text{Ker } \mathbf{L}_\alpha. \tag{5.41}$$

Hence, this subspace is defined as

$$\mathbf{S}^* = \begin{bmatrix} \mathbf{u}^*_{\mathbf{M_1}} & \mathbf{u}^*_{\mathbf{M_2}} \end{bmatrix} = \begin{bmatrix} v_{x1} & v_{x2} \\ v_{y1} & v_{y2} \\ \omega_{z1} & \omega_{z2} \end{bmatrix}. \tag{5.42}$$

which corresponds analytically to

$$\mathbf{S}^* = \begin{bmatrix} \cos\alpha & \ell\sin\alpha \\ \sin\alpha & -\ell\cos\alpha \\ 0 & 1 \end{bmatrix}. \tag{5.43}$$

The first column of $\mathbf{S}^*$ induces a translation motion along the sound source direction, that is along $\overrightarrow{\mathbf{MX_s}}$. The second column describes a rotation around $\overrightarrow{z_M}$ axis combined with a translation. This last motion can be geometrically represented by a circle of radius ℓ centred on $\mathbf{X_s}$. Hence, similarly to the ILD case, such task consists principally in orienting the robot with respect to the location of the source.

Thus, any linear combination of these two motions implies infinite poses to complete the task $\alpha = \alpha^*$ (see Figure 5.3), from which we set up a first proposition.

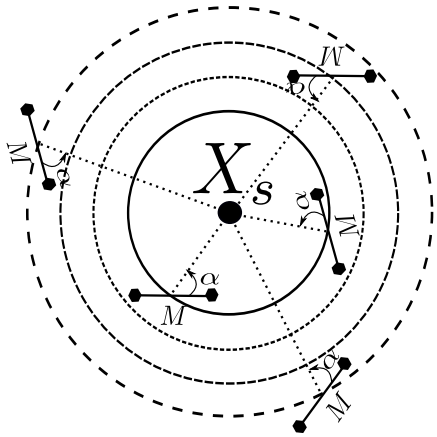


Figure 5.3: With one sound source several poses exists for a given positioning task expressed by $\alpha = \alpha^*$. These poses lie on circles of radius ℓ centred on the sound source location.

Proposition 1 *For each random position of the microphones defined by* $\mathbf{M}$*, there exists one orientation* $\theta_{\mathbf{M}}$ *of the microphones such that* $\alpha = \alpha^*$ *.*

5.2.5 Stability analysis

In this part, we analyze the stability properties of our sensor-based control through the interaction matrix modelled in (5.34) and (5.29). The stability properties tailor the approximation of the unknown parameters in the feature Jacobian matrix. With the same reasoning developed for ILD context, in the previous chapter, the Lyapunov global asymptotic stability condition is obtained when

$$\mathbf{L}_\tau \widehat{\mathbf{L}_\tau^+} > 0, \tag{5.44}$$

knowing that $\widehat{\mathbf{L}_\tau} \in \mathbb{R}^{1\times 3}$. In order to determine the stability conditions, let us focus first on the Jacobian matrix related to the DOA α. In this case, the stability is obtained when

$$\mathbf{L}_\alpha \widehat{\mathbf{L}_\alpha^+} > 0. \tag{5.45}$$

From (5.28) and (5.29), we have

$$\mathbf{L}_\alpha \widehat{\mathbf{L}_\alpha^+} = \begin{bmatrix} \frac{\sin\alpha}{\ell} & -\frac{\cos\alpha}{\ell} & -1 \end{bmatrix} \begin{bmatrix} \frac{\widehat{\ell}\sin\alpha}{1+\widehat{\ell}^2} \\ -\frac{\widehat{\ell}\cos\alpha}{1+\widehat{\ell}^2} \\ -\frac{\widehat{\ell}^2}{1+\widehat{\ell}^2} \end{bmatrix}. \tag{5.46}$$

Equation (5.46) can be simplified to

$$\mathbf{L}_\alpha \widehat{\mathbf{L}_\alpha^+} = \frac{\widehat{\ell} + \ell\widehat{\ell}^2}{\ell(1+\widehat{\ell}^2)}. \tag{5.47}$$

Thus, it appears that for ensuring the stability condition $\mathbf{L}_\alpha \widehat{\mathbf{L}_\alpha^+} > 0$, it is sufficient to set $\widehat{\ell} > 0$. However, this demonstration is valid under the condition that the DOA injected in $\widehat{\mathbf{L}_\alpha}$ is not too coarse. Indeed if we consider an approximation $\widehat{\alpha}$ estimated from τ, (5.47) is more exactly written as

$$\mathbf{L}_\alpha \widehat{\mathbf{L}_\alpha^+} = \frac{\widehat{\ell} \sin\alpha \sin\widehat{\alpha} + \widehat{\ell}\cos\alpha\cos\widehat{\alpha} + \ell\widehat{\ell}^2}{\ell(1+\widehat{\ell}^2)}. \tag{5.48}$$

Considering that $\ell > 0$ and $\widehat{\ell} > 0$, the stability is ensured as soon as $sign(\cos\widehat{\alpha}) = sign(\cos\alpha)$ and $sign(\sin\widehat{\alpha}) = sign(\sin\alpha)$. This interesting result confirms that even with rough approximation of α, the controller should be able to converge towards the desired configuration. Generally, any technique of ITDs estimation (referenced in Chapter 2.2), and by extension DOA estimation, provides sufficiently accurate results so that $sign(\cos\widehat{\alpha}) = sign(\cos\alpha)$. The second condition $sign(\sin\widehat{\alpha}) = sign(\sin\alpha)$ is as expected related to the *front-back ambiguity* inherent to ITDs. As a consequence, under the far-field assumption, the stability of the control scheme based on $\widehat{\mathbf{L}_{\tau_r}}$ in (5.29) is ensured as soon as $\widehat{\ell} > 0$, the DOA α and then the ITD τ is not too coarse (see (5.18) and Figure 5.2) and the *front-back ambiguity* is solved.

For generalized stability properties (*i.e.*, valid in the near-field), the Lyapunov stability should also be evaluated on (5.34). First, to ease the comprehension and the reading of the analysis, let us denote the interaction matrix $\widehat{\mathbf{L}_\tau}$ as

$$\widehat{\mathbf{L}_{\tau_r}} = \begin{bmatrix} \frac{\widehat{N}_1}{\widehat{\ell}_1\widehat{\ell}_2} & \frac{\widehat{N}_2}{\widehat{\ell}_1\widehat{\ell}_2} & \frac{\widehat{N}_3}{\widehat{\ell}_1\widehat{\ell}_2} \end{bmatrix} \tag{5.49}$$

where $\widehat{N_1} = \widehat{x}_s\tau - \frac{A}{2}(\widehat{\ell}_1 + \widehat{\ell}_2)$, $\widehat{N_2} = \widehat{y}_s\tau$ and $\widehat{N_3} = \frac{A}{2}(\widehat{\ell}_1 + \widehat{\ell}_2)\widehat{y}_s$. Similarly $\mathbf{L}_\tau$ can be defined with N_1, N_2, N_3, ℓ_1 and ℓ_2. From this convention of writing, (5.44) becomes

$$\mathbf{L}_{\tau_r}\widehat{\mathbf{L}_{\tau_r}^+} = \frac{\widehat{\ell}_1\widehat{\ell}_2}{\ell_1\ell_2}\frac{N_1\widehat{N_1} + N_2\widehat{N_2} + N_3\widehat{N_3}}{\widehat{N_1}^2 + \widehat{N_2}^2 + \widehat{N_3}^2}. \tag{5.50}$$

Considering of course $\ell_i > 0$ and $\widehat{\ell}_i > 0$, the Lyapunov stability is ensured as soon as

$$N_1\widehat{N}_1 + N_2\widehat{N}_2 + N_3\widehat{N}_3 > 0. \tag{5.51}$$

The latter equation can be developed as:

$$y_s\widehat{y}_s\tau^2 + \frac{A^2}{4}y_s\widehat{y}_s(\widehat{\ell}_1+\widehat{\ell}_2)(\ell_1+\ell_2) + (x_s\tau - \frac{A}{2}(\ell_1+\ell_2))(\widehat{x}_s\tau - \frac{A}{2}(\widehat{\ell}_1+\widehat{\ell}_2)) > 0. \tag{5.52}$$

Hence, it can be deduced that a sufficient condition of stability is obtained with $sign(\widehat{x}_s) = sign(x_s)$ and $sign(\widehat{y}_s) = sign(y_s)$. Indeed (5.52) is always positive for these conditions, since $(\widehat{x}_s\tau - \frac{A}{2}(\widehat{\ell}_1 + \widehat{\ell}_2)) < 0$ and $(x_s\tau - \frac{A}{2}(\ell_1 + \ell_2)) < 0$ owing to $\frac{\ell_1+\ell_2}{2} > x_s$ and $|\tau| < A$ (see (5.10)).

In the end, the stability conditions of the controller, assuming the far-field approximation or the proper modelling of the ITD are exactly the same. Indeed, from

(5.2) the conditions $sign(\widehat{x}_s) = sign(x_s)$ and $sign(\widehat{y}_s) = sign(y_s)$ can be rewritten as:

$$\begin{cases} sign(\widehat{\ell}\cos\widehat{\alpha}) = sign(\ell\cos\alpha) \\ sign(\widehat{\ell}\sin\widehat{\alpha}) = sign(\ell\sin\alpha) \end{cases} \tag{5.53}$$

which amounts to $\widehat{\ell} > 0$, $sign(\cos\widehat{\alpha}) = sign(\cos\alpha)$ and $sign(\sin\widehat{\alpha}) = sign(\sin\alpha)$, the conditions described for the control scheme based on the far-field assumption. Both modelling of the ITD stresses the necessity to disambiguate the source location on the hyperbola $\mathcal{H}$ though. The conditions $sign(\widehat{x}_s) = sign(x_s)$ and $sign(\widehat{y}_s) = sign(y_s)$ are required. The first condition is easily fulfilled from the sign of τ as discussed above in (5.13). However, the second condition related to the *front-back ambiguity* cannot be ensured from the ITD measurement. Nonetheless, experimentally, it may be possible to leave the instability zone (*i.e.*, $sign(\widehat{y}_s) \neq sign(y_s)$) to reach a pose where all the above conditions are completed, as already demonstrated in Chapter 4.2.5. Eventually, the conditions resulting from the current study underline infinite solutions ensuring global asymptotically stability, for both cases of ITD modelling. From these results, one can notice the strong similarity between the stability properties for the ITD and for the ILD. Indeed, the stability conditions are exactly the same for both cases, which is not surprising since these cues carry the same type of information: the azimuth of the source.

5.2.6 Experimental results

5.2.6.1 Distance of observation evaluation

Before experimenting the proposed control scheme, it is essential to evaluate the difference between the interaction matrix based on the far-field assumption $\widehat{\mathbf{L}_{\tau_f}}$ and $\widehat{\mathbf{L}_{\tau_r}}$ extracted from the actual geometry of the ITD. We then tested these two models through simulations. The simulation environment designed from *Roomsimove* [BOV12] consists in a room of 8×6 m^2 in which a positioning task is performed with respect to a speech sound $\mathbf{X}_s$. The positions of $\mathbf{X}_s$ are defined so that α varied by $18°$ at given distance ℓ within the range $[18°, 162°]$ as depicted in Figure 5.4. When considering one sound source, since the task mainly consists in orienting the robot towards a direction, we considered only rotational velocity of the control that lead to

$$\omega_{z_f} = -\lambda\nu(\tau - \tau^*) \tag{5.54}$$

where ω_{z_f} is the control based on the far-field assumption. Similarly the control ensued from the rigorous geometrical definition of ITDs ω_{z_f} is defined as:

$$\omega_{z_r} = -\lambda\frac{\frac{A}{2}(\widehat{\ell}_1 + \widehat{\ell}_2)\widehat{y}_s}{\widehat{\ell}_1\widehat{\ell}_2}(\tau - \tau^*) \tag{5.55}$$

In (5.54), the parameters to be estimated are set to $\widehat{x}_s = -1 \times sign(\tau)$ and $\widehat{y}_s = 1$. $\widehat{\ell}_1$ and $\widehat{\ell}_2$ can afterwards be deduced by using (5.1). The task consists in facing the sound source, then $\tau^* = 1$. The ITDs τ are estimated in real-time, for both configuration, by using GCC-PHAT algorithm (see Chapter 2.2.3). Eventually we

considered anechoic conditions, since these simulations aim principally to evaluate the accuracy of these two modelling with respect to the distance of observation of the sound. The orienting task was repeated from far-field and near-field distances. More explicitly ℓ was set respectively at 0.5 m of the microphones in the near-field and 2 m for the far-field with $d = 0.3$ m. In the results illustrated in Table 5.1, we also compared the modelling to an open loop localization approach, in which the accuracy of the azimuth is evaluated from AEG model (see Chapter 2.2.2). The absolute orientation error (azimuth error in the localization case) in degree are reported.

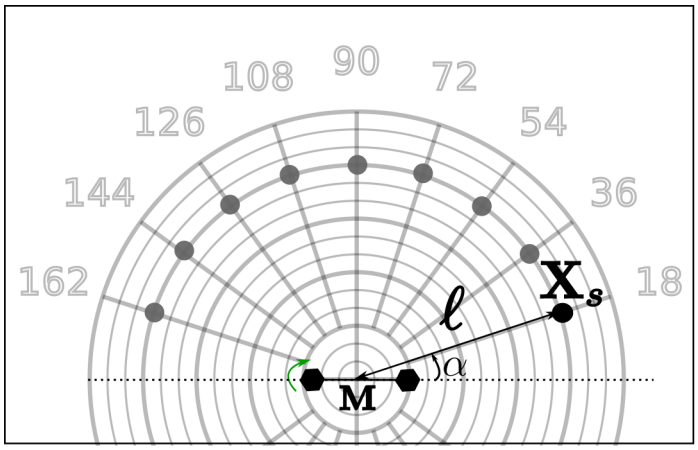


Figure 5.4: A simulated task that consists in facing a sound source ($\tau^* = 1$) in a $8 \times 6\ m^2$ room. The sound source positions are varied so that $\alpha \in [18°, 162°]$ by step of 18° and $\ell \in \{0.5, 2\}$ m for a near-field and far-field observation distance.

		α (degrees)								
		18	**36**	**54**	**72**	**90**	**108**	**126**	**144**	**162**
Near-field	$\mathbf{u_f}$	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
	$\mathbf{u_r}$	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
	loc	7.1	8.2	6.3	3.9	0	3.9	6.8	8.2	7.1
Far-field	$\mathbf{u_f}$	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
	$\mathbf{u_r}$	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1	< 1
	loc	< 1	1.1	< 1	< 1	< 1	< 1	< 1	1.1	< 1

Table 5.1: Absolute error of the orientation task described in Figure 5.4 considering a control scheme using on far-field assumption ω_{z_f} or a control scheme without assumption ω_{z_r}. These two methods are compared to a baseline approach based on localization **loc**.

The simulations show that both control schemes ω_{z_r} and ω_{z_f} give excellent results independently of the distance of observation. This result could be expected from the stability conditions that are exactly the same for the two approaches. From the stability perspective, the actual interaction matrix related to the far-field assumption $\mathbf{L}_{\tau_f}$ is just an approximation of $\mathbf{L}_{\tau_r}$ so that it can be written

$$\mathbf{L}_{\tau_f} = \widehat{\mathbf{L}_{\tau_r}}. \tag{5.56}$$

The values ℓ (and τ) are directly reflected in $\widehat{\mathbf{L}_{\tau_r}}$ through $\widehat{x}_s$, $\widehat{y}_s$ and $\widehat{\ell}_i$. Hence, without violating the stability conditions of $\mathbf{L}_{\tau_r}$, it is ensured that a control scheme based on $\mathbf{L}_{\tau_f}$ will converge in the near-field. This outcome is also valid for $\widehat{\mathbf{L}_{\tau_f}}$ under satisfying stability conditions. Consequently the ITD-based positioning task can be performed accurately with a far-field assumption in the near-field. Our approach is independent of the distance of observation. Eventually, the results exhibited by our approach can also be compared to a more conventional positioning task based on localization. Conversely to our method, in the near-field, the accuracy of the localization is particularly degraded, because of the far-field assumption. This result corroborates the geometrical modelling of the ITD detailed in Section 5.1.2, in which the hyperbola $\mathcal{H}$ does not fit with its asymptotes in the near-field. Although the simulation results of localization remains acceptable in the near-field (in anechoic conditions), when facing real environments the errors may be accentuated by reverberation, noise and other perturbations.

As a direct consequence of these results, we can already emphasize the potential robustness of our control scheme towards approximations or inaccurate estimations unlike open-loop approaches based on localization. In the following, all the experiments and analysis are based on $\widehat{\mathbf{L}_{\tau_f}}$, that has a more convenient form. For more simplicity, we rename $\mathbf{L}_{\tau_f}$ as $\mathbf{L}_\tau$.

5.2.6.2 Preliminaries: robot modelling and control scheme

To apply the proposed framework, the same system as in Chapter 4.2.5 is modelled: a non-holonomic unicycle robot endowed with two microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ as illustrated on Figure 5.5. In this context, we aim to control the two DOFs of the robot characterized by the control input $\dot{\mathbf{q}}$ is given by (u, ω), respectively the translation and rotation velocities. In addition to $\mathcal{F}_m$ related to the microphones, the frame $\mathcal{F}_r(\overrightarrow{x_R}, \overrightarrow{y_R})$ is attached to the robot. D_x denotes the distance between the center of the robot $\mathbf{R}$ and the midpoint of the microphones $\mathbf{M}$.

From this modelling, the control input $\dot{\mathbf{q}}$ can be computed from the following relationship:

$$\dot{\tau} = \mathbf{J}_\tau \dot{\mathbf{q}} \tag{5.57}$$

where $\mathbf{J}_\tau$ is the Jacobian feature matrix introduced in Chapter 3.2.1 that is equal to

$$\mathbf{J}_\tau = \mathbf{L}_\tau \mathbf{J_r}. \tag{5.58}$$

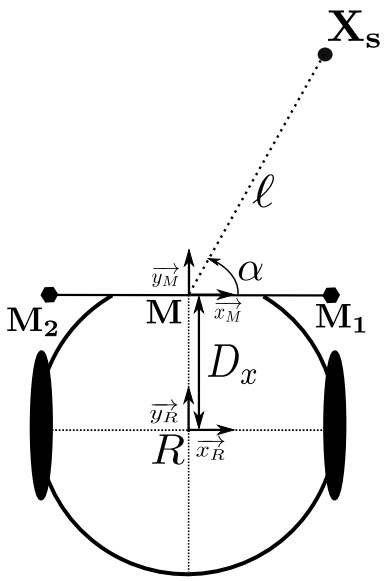


Figure 5.5: Modelling of the robotic platform

$\mathbf{J}_r$ is the robot Jacobian obtained from Figure 5.5 as

$$\mathbf{J_r} = \begin{bmatrix} 0 & D_x \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{5.59}$$

Hence, the control scheme of the robot is

$$\dot{\mathbf{q}} = -\lambda \widehat{\mathbf{J}_\tau^+}(\tau - \tau^*). \tag{5.60}$$

As mentioned in the task analysis, a control input characterized by the angular velocity ω only is sufficient to achieve any task involving one sound source. In that particular case where u would always be equal to 0 and the Jacobian matrix reduces to

$$\widehat{\mathbf{J}_\tau} = \frac{A\widehat{\ell}\nu - D_x\nu^2}{A\widehat{\ell}}. \tag{5.61}$$

The control input becomes

$$\dot{q} = -\lambda \frac{1}{\nu\left(1 - \frac{D_x}{A\widehat{\ell}}\nu\right)}(\tau - \tau^*) \tag{5.62}$$

Excluding the degenerate case where $\alpha = 0$ or $\alpha = 180$ (i.e., $\nu = 0$), singularities of the control scheme could in principle occur if $\widehat{\ell} < D_x$. Hopefully, this is impossible in practice. Indeed, knowing that $0 \leq \nu \leq A$, the denominator of (5.62) can never vanish as soon as $\widehat{\ell} > D_x$. In the same way, the global asymptotic stability of the system is satisfied as soon as $\ell > D_x$ and $\widehat{\ell} > D_x$.

When the two DOF of the robot (u, ω) are controlled, the Jacobian matrix is given by

$$\widehat{\mathbf{J}_\tau} = \begin{bmatrix} \frac{\tau\nu}{A\widehat{\ell}} & \frac{A\widehat{\ell}\nu - D_x\nu^2}{A\widehat{\ell}} \end{bmatrix} \tag{5.63}$$

and the control input is obtained as:

$$\dot{q} = -\lambda \frac{\tau^2 + (\nu D_x - A\widehat{\ell})^2}{A^2\widehat{\ell}^2} \begin{bmatrix} \frac{\tau\nu}{A} \\ -\nu(\frac{\nu D_x}{A} - \widehat{\ell}) \end{bmatrix} (\tau - \tau^*) \tag{5.64}$$

The control scheme input is not singular as soon as $\widehat{\ell} > 0$, which corresponds to the Lyapunov stability condition already demonstrated previously.

5.2.6.3 ITD estimation and tracking

Prior to the experimental validations, it is essential to characterize the sound processing and more particularly ITDs estimation. Indeed unlike ILDs that can be computed uniquely by integrating the sound signal over a particular time window, ITDs are estimated from a more complex process. Recalling the methods of ITDs estimation detailed in Chapter 2.2.3, from GCC-PHAT [KC76], the ITD τ can be obtained by comparing the temporal signals $x_1(t)$ from the microphone $\mathbf{M_1}$ and $x_2(t)$ from $\mathbf{M_2}$ in the spectral domain with:

$$\boldsymbol{R}_{1,2}(\tau) = \sum_{f}^{F} \frac{\phi_{x_1,x_2}(f)}{|\phi_{x_1,x_2}(f)|} e^{\varphi(\tau)}. \tag{5.65}$$

The cross-spectral power density ϕ_{x_1,x_2} is defined in our case by

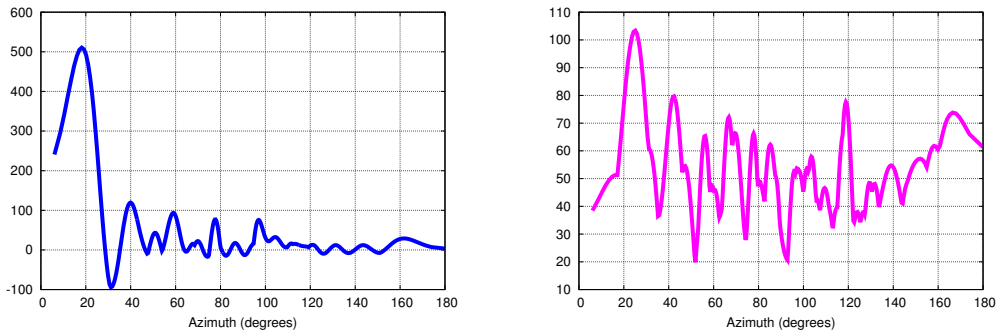
$$\phi_{x_1,x_2}(f) = \max_{l} \; X_1(f,l) X_2^*(f,l). \tag{5.66}$$

$X_1(f,l)$ and $X_2^*(f,l)$ are respectively the Fourier transform of $x_1(t)$ and the conjugate of the Fourier transform of $x_2(t)$. The maximum peak of the GCC-PHAT function gives an estimation of the actual ITD and can therefore be written as:

$$\widehat{\tau} = \underset{\tau}{\operatorname{argmax}} \; \boldsymbol{R}_{1,2}(\tau) \tag{5.67}$$

with $\tau \in [-A, A]$ corresponding to a sound direction of arrival α from 0° to 180°. In ideal conditions, the maximum peak of the latter function should correspond to the ITD of the sound source as exposed in Figure 5.6a.

In real world conditions, the result of the GCC-PHAT function is more contrasted. Instead of having one dominating peak, which makes immediate the estimation of the ITD, several plausible but spurious peaks appear because of the reverberation and noise (see Figure 5.6b). These conditions may also alter the actual peak related to the source that would be less sharp and accurate. Fortunately, this problem can be handled by our control framework that is robust to approximate ITDs. Therefore, under this conditions, p peaks ($p > 1$) may be considered among which the correct peak should be found. Furthermore, when considering specific signals such



(a) (b)

Figure 5.6: Output of the GCC-PHAT function under anechoic conditions (a) and reverberant conditions (b).

as speech, it is very likely to record sparse data (temporally and spectrally). Not all the frames processed contain relevant information, since speech is a non-stationary and intermittent signal.

This characteristic emphasizes the strong necessity of a tracking algorithm to process ITDs. Independently of the intrinsic value of the features, AS, as many other sensor-based approaches, is built upon a correct tracking of the input features. Tracking the ITDs corresponding to the actual source among a set of observations may be an issue for the correct achievement of the desired task. If ITDs that do not correspond to the source are tracked the positioning task would certainly fail. In robot audition literature, the problem of tracking the ITDs is quite new, with the recent interest in *active audition* and moving source/robot as stated in Chapter 1.2. The common approach consists in probabilistic techniques such as the particle filter used in [VMR07] and [MP10]. These two approaches are complemented with a speech activity detection to account for periods of silence. In [EMN+15] a probability hypothesis density tracker is proposed in order to skip the activity detection. The aforementioned approaches are relying on array-based approach for more robustness. In binaural context [PDA12] and [NCVC16] are based on a mixture of Kalman filters. However these methods have only been partially validated in real environments [PBD+14]. Being able to accurately track dynamic ITDs is still a challenging task in robot audition.

Yet, AS can provide added value to this tracking problem. One of the benefit of coupling motion and perception, as it is in AS, lies in the prediction that limits the scope of erroneous measurements. More specifically, the Jacobian matrix $\mathbf{J}_\tau$ can be used to infer the evolution of the tracked ITDs in the next time frame. Given τ as the state x and the velocity $\dot{\mathbf{q}}$ applied to the robot, a local prediction model based on the Jacobian matrix (5.58) is simply given as follow:

$$\begin{cases} \dot{x}(k) = \widehat{\mathbf{J}_\tau}\dot{\mathbf{q}} \\ x(k+1) = x(k) + T_e\dot{x}(k) \end{cases} \tag{5.68}$$

in which T_e refers to the sampling time of the control loop. Nonetheless the predicted

τ is not as accurate as the genuine ITD, because of the approximation $\widehat{\mathbf{J}}_\tau$ used in (5.60) in which $\ell_i = \widehat{\ell_i}$. Several methods such as state space observer [DLORG08] could be used to obtain a better estimation of $\widehat{\ell_i}$, but the closed-loop control scheme is sufficiently robust to cope with a rough approximation of ℓ_i. Despite this rough approximation, the prediction step gives useful information that can also be used when the source is not active. If no ITD measurements "fit" with the predicted value, it is very likely that the source is inactive. Hence, the unavailable ITD could be replaced by this prediction. Without complex tracking methods, it should then be possible to track the correct ITDs in realistic environments.

In our application, the tracking procedure is divided into two steps. In the first step, the goal is to find the correct ITD in respect of the number of active sound sources defined beforehand. This step assumes that the robot and the sound source(s) are not moving. Then the set of peaks obtained from the GCC function is observed during a given number of time frames. A clustering algorithm based on the Euclidean distance between the ITDs is then applied to detect the frequency of appearance of a given ITDs τ_i. By combining this frequency to the mean appearance rank of each τ_i, the most probable ITD is then selected. This method is applied to the initial pose to retrieve $\tau(t_0)$ but could also be applied to retrieve each τ^* when the desired pose is not characterized by an obvious ITD. For some simple cases, each τ^* is defined manually, for instance for a task that consists in orienting the robot towards the sound source (*e.g.*, $\tau^* = 0$). The second step is used during the motion of the robot, and consists in finding the genuine ITD among the set of observations given by the GCC-PHAT function. Knowing the previous value of the ITD, we simply select the closest peak in the current frame, by taking into account our predicted ITD.

It is clear that the tracking method proposed is not optimal. More evolved solutions could complement our approach with the methods available in the literature (particle filtering, probabilistic data association...). They would, however, increase the computational cost of the control scheme. We have preferred the simple and efficient solution exposed above in order to stress the added value of our approach, and more particularly the robustness of the control scheme to inaccurate and/or punctual errors.

5.2.6.4 Experimental setup

In order to further validate the effectiveness of our approach we conducted experiments on the *Pioneer 3DX* robot. Two microphones connected to a sound card 8SoundsUSB [AC+] were used. The sound card operates at a frequency of 48 kHz, and provides windows frames of 256 samples. The ITD is computed from 10 consecutive windows frames (e.g 50 ms), that are sub-sampled at 16 kHz. Processing the sound signal at a frequency of 16 kHz gives two advantages: better results are obtained from uttered speeches that is less sparse at low frequencies and the processing time is reduced with less samples to analyze. Consequently, the global control framerate is around 12 Hz. The tests were conducted in a room with a reverberation time $\mathrm{RT}_{60} \approx 580$ ms. Moreover, the measured SNR is around 20 dB in presence of typical noise such as computer noise and ventilation in the room. The parameters

d	0.31 m
c	343 m.s^{-1}
$\widehat{\ell}$	1 m
A	$0,00090379$ s
$\lambda(x)$	$5e^{(-4000x)}$

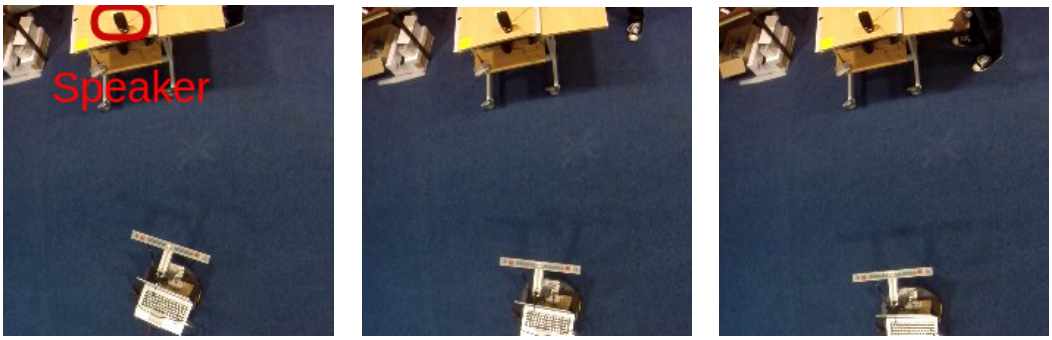
Table 5.2: Experimental settings

used for the experiments are given in Table 5.2. An adaptive gain $\lambda(x)$ in which x refers to the norm of the error e is used to smooth the robot motion.

5.2.6.5 Typical ITD-based positioning task

In this experiment we considered one sound source that corresponds to a female voice recording of 10 s played in loop. The speech pauses are not removed from the sound signal. The task to be performed by the robot is to face the sound source, τ^* is set to 0. The two DOF are controlled from the control input given by (5.64). The experiment can be divided into two steps.

In the first step the robot is randomly oriented with respect to the speaker. Nonetheless this orientation is set so that the source stays in the front-side of the robot. In this part of the experiment the robot correctly positioned itself in the direction of the static sound source, as illustrated by the Figure 5.7.

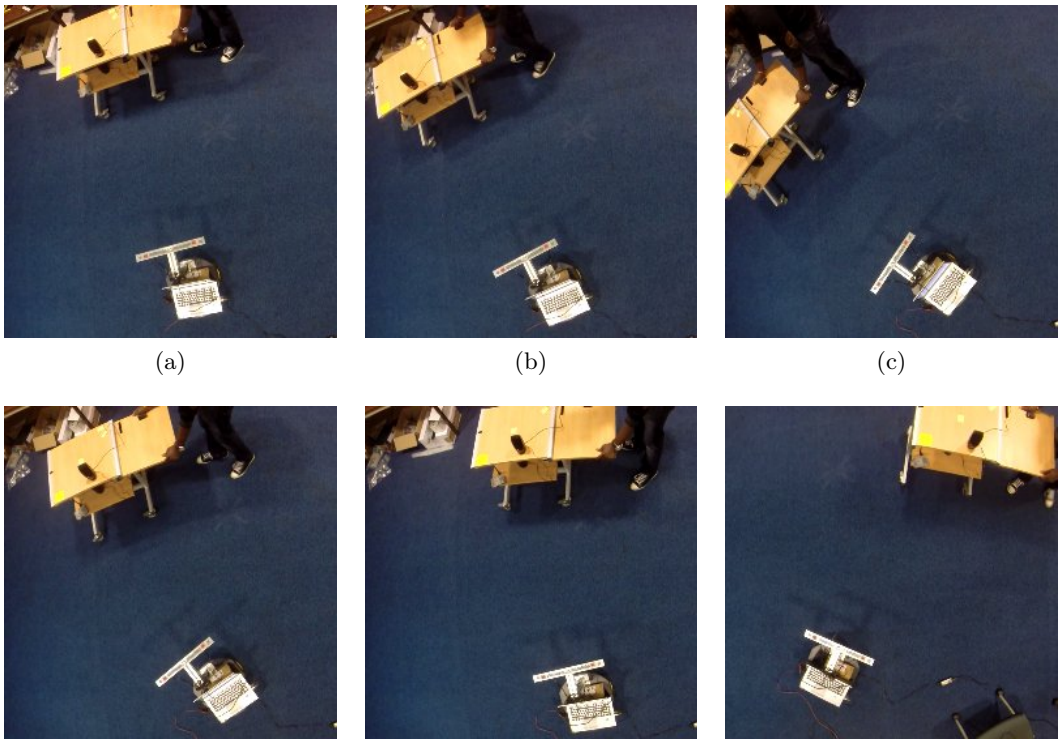


(a) (b) (c)

Figure 5.7: First phase of the experiment: the robot accurately orients towards the sound source

Subsequently the sound source was moved from one side in the environment and at different distances. As a result, the robot constantly moved in the direction of the source (see Figure 5.8). The robot accurate behaviour is supported by the data exposed in Figure 5.9, that are extracted from the experiment. During the first 5 s, the error follows an exponential decrease pattern as modelled in the control scheme.

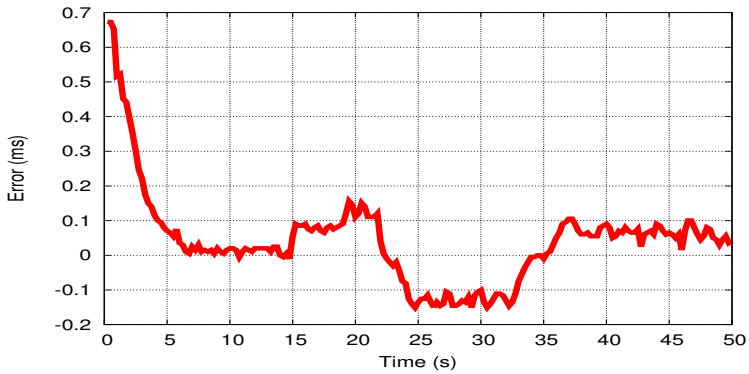
In the following, when the source moves, the error is compensated by the motion of the robot. As already mentioned in the previous chapter, in ILD case, there is still a small tracking error (response delay), that is illustrated by the error curve that do not stay at 0. Nonetheless, the task is correctly achieved since the lag between the robot orientation and the desired one did not exceed 10° despite the low dynamic response of the mobile robot. Knowing that no specific information about the source motion is used in the control scheme, the behaviour of the robot remains satisfactory. Furthermore, this experiment acknowledges the relevance of the tracking method. We can notice in Figure 5.9c several false measurements represented by the green dots while the actual ITD is accurately tracked. These spurious measurements may be caused by reverberation, or simply noise. The effect of the noise can also be noticed by the aligned green dots at $\alpha = 90°$ during all the frames of the task. This diffuse noise is certainly caused by the ventilation system in the room. It can be noted that no particular signal enhancement has been used for this experiment. Unlike the experiments proposed in the previous chapter, we are not attached here to solve the *front-back ambiguity*. The singularities related to the ITD hyperbola (*i.e.*, $\tau \pm A$ or $\alpha = 0°$, $\alpha = 180°$) would make the control scheme unstable.



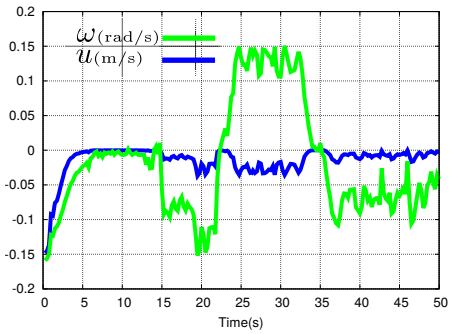
(a) (b) (c)

(d) (e) (f)

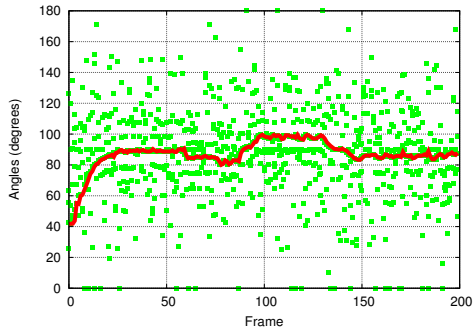
Figure 5.8: Second phase of the experiment: the robot is accurately tracking a sound source that is moving



(a) Error $\tau - \tau^*$



(b) Control input



(c) Observed (green dots) and tracked (red line) α deducted from ITDs

Figure 5.9: Experimental result of the ITD-based task illustrated in Figures 5.7 and 5.8

5.2.7 Numerical evaluation

The proposed framework has also been evaluated in terms of reverberation and noise in a numerical simulation. An acoustic room of 5 by 5 m^2 is simulated, in which a sound source corresponding to the same female speech as the experiments in loop is played. Each ITD is extracted from a frame of 100 ms. This frame size was chosen for real-time purpose. The experiment is repeated several times with 12 source locations so that the azimuth angle at the starting pose of the microphones varies between 30° and 150° with a resolution of 10° (the azimuth angle 90° is not considered). The goal of the task is to orient the robot towards the sound source (i.e $\tau^* = 0$ or $\alpha = 90°$). The correct achievement of the tasks as well as the erroneous and missing ITD are evaluated. A measurement is considered erroneous if the error between the actual ITD and the estimated ITD leads to an error of 5° or more in the corresponding DOA α. A task is said to be completed when $\tau = \tau^*$ at the end of the servo.

Robustness to noise and reverberation

Table 5.3 sums up the results of the simulation. As expected the performance is decreased in presence of high level of noise coupled with reverberation. Under 20

SNR(dB)	RT_{60} (s)				Total
	0	0.05	0.1	0.2	
30	1	1	1	1	1
25	1	1	1	1	1
20	1	0,916	0,667	0	0,646
15	1	0,667	0,416	0,083	0,541
10	0,667	0,167	0	0	0,208
Total	0,944	0,625	0,514	0,347	

Table 5.3: Rate of achieved tasks considering several reverberation and noise conditions

dB of noise, the task is affected by the reverberation. However, most of the time, the robot did not reach the correct pose because of a wrong initialization of the ITD measurement (e.g $\alpha(t = 0) = 90°$). These failures stress the importance of being able to track the correct ITD and more importantly to initialize correctly the tracking algorithm. Hence, a more elaborated initialization step could improve the rate of completion of the task for low SNR. Globally, we remark that this approach is more sensitive to high level of noise than reverberation. This result fits with the method chosen for the ITD calculation (GCC-PHAT), known to be less robust to noise. The PHAT processor whitens the signal for all frequencies without any selectivity of frames containing noise or speech. Approaches such as RWPHAT (see Chapter 2.2.3) would undeniably improve these results. On the other hand, in a context suitable for GCC-PHAT, that is to say with a high SNR, the positioning task is accurately performed. A more thorough study of these results confirms that the control scheme is able to cope with erroneous/missing measurements.

Robustness to erroneous measurements

Among the achieved tasks, we focus now on the rate of erroneous/missing ITDs. The Figure 5.10 illustrates this rate for a noise level of 30 and 25 dB. Lower SNR are not considered in this figure since the different number of achieved tasks would bias the global analysis. Although these tasks were all achieved correctly, several erroneous or missing measurements are faced by the control scheme mainly because of the speech pauses. As expected, this rate of erroneous ITD increases logically with higher level of noise and reverberation. But the control scheme is still able to complete the task by using the prediction derived from the interaction matrix, even for cases where nearly 20% of the measurements are missing or erroneous. This result show the effectiveness of our method to cope with punctual erroneous measurements.

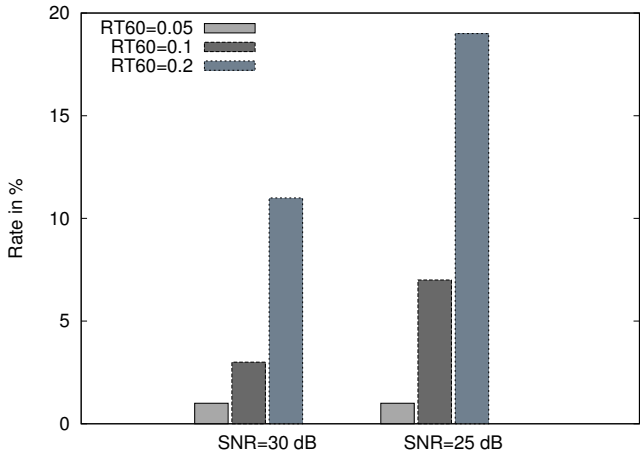


Figure 5.10: Average rate of missing/erroneous ITD considered for correctly achieved tasks for 30 and 25 dB of SNR

5.3 Multi-source tasks

In the second part of this chapter, we address the scenarios where several sound sources are active in the scene. Here, the idea is to define a positioning task with respect to the ITDs extracted from each sound source. Under the far-field assumption that relates each ITD τ_i to the DOA α_i such positioning task can be referred to the field of beaconing or bearing-only homing/navigation. This kind of approach is bio-inspired and such process can be observed in nature through the localization system of bees [CC87]. One of the first application about bearing-based robot control can be found in [LDW91] where the navigation of a mobile robot is governed by the beacon extracted by an array of sonar transducers. In the field of vision-based control several works also exploit beacons. For instance, [Cor03] proposes a planar visual servoing navigation system based on these features. This work inspired more recently [LPCS10] that develops a homing system based only on bearing measurements, without knowledge of the distance to the visual landmarks. This context is also exploited nowadays by the field of multi-robot localization [MPS05] or formation control of ground robots [FM02] or aerial vehicles [FMG+12] that is essentially based on bearing measurements. Such aforementioned tasks could also be performed by exploiting sound.

5.3.1 Case of two sound sources

5.3.1.1 Control scheme and task analysis

We consider now the configuration where two sound sources $\mathbf{X_{s_1}}$ and $\mathbf{X_{s_2}}$ are present in the acoustic scene. In this configuration the interaction matrix is obtained by

merely stacking (5.25) for each sound source:

$$\mathbf{L}_{\boldsymbol{\alpha}} = \begin{bmatrix} \frac{\sin\alpha_1}{\ell_1} & -\frac{\cos\alpha_1}{\ell_1} & -1 \\ \frac{\sin\alpha_2}{\ell_2} & -\frac{\cos\alpha_2}{\ell_2} & -1 \end{bmatrix}, \tag{5.69}$$

which is equivalent to

$$\mathbf{L}_{\boldsymbol{\tau}} = \begin{bmatrix} -\frac{\nu_1^2}{A\ell_1} & \frac{\tau_1\nu_1}{A\ell_1} & \nu_1 \\ -\frac{\nu_2^2}{A\ell_2} & \frac{\tau_2\nu_2}{A\ell_2} & \nu_2 \end{bmatrix} \tag{5.70}$$

with the ITDs τ_1 and τ_2 respectively related to $\mathbf{X_{s_1}}$ and $\mathbf{X_{s_2}}$, and $\nu_i = \sqrt{A^2 - \tau_i^2}$. Similarly to the single-source task, we consider a control scheme based on the approximation of the latter interaction $\widehat{\mathbf{L}_{\boldsymbol{\tau}}}$.

Once again, the task related to the use of such control scheme can be characterized from the virtual linkages approach applied on (5.69). This time $\mathbf{S}^* \in \mathbb{R}^{3\times 1}$ implies a class 1 virtual link:

$$\mathbf{S}^* = \begin{bmatrix} \ell_1\cos\alpha_2 - \ell_2\cos\alpha_1 \\ \ell_1\sin\alpha_2 - \ell_2\sin\alpha_1 \\ \sin(\alpha_1 - \alpha_2) \end{bmatrix}. \tag{5.71}$$

Geometrically ${\mathbf{S}_2}^*$ refers to the circumscribed arc of circle characterized by $\mathbf{X_{s_1}}$, $\mathbf{X_{s_2}}$ and $\mathbf{M}$, as it can be proved by the inscribed angle theorem. Let δ be $\angle\mathbf{X_{s_1}}\mathbf{M}^*\mathbf{X_{s_2}}$ (i.e $\delta = \alpha_2^* - \alpha_1^*$) from a specified pose as it appears in Figure 5.11a. This theorem exposes that given $\mathbf{X_{s_1}}\mathbf{X_{s_2}}$, the set of points $\mathbf{M}$ in the plane for which the angle $\angle\mathbf{X_{s_1}}\mathbf{M}\mathbf{X_{s_2}}$ is equal to δ is an arc belonging to the circumscribed circle of $\mathbf{M}^*\mathbf{X_{s_1}}\mathbf{X_{s_2}}$. Then by considering a random point $\mathbf{M}$ on this arc, from Proposition 1, it is guaranteed to find an orientation $\theta_{\mathbf{M}}$ such that $\alpha_1 = \alpha_1^*$ (or $\alpha_2 = \alpha_2^*$). Knowing that δ remains constant, $\alpha_2 = \alpha_2^*$ (or $\alpha_1 = \alpha_1^*$) is guaranteed from $\delta = \alpha_2^* - \alpha_1^*$ This result sets up the following proposition:

Proposition 2 *For each random position* $\mathbf{M}$ *on the arc of the circumscribed circle defined by* $\mathbf{X_{s_1}}$, $\mathbf{X_{s_2}}$ *and* $\mathbf{M}^*$, *there exists one orientation* $\theta_{\mathbf{M}}$ *of the microphones so that* $\alpha_1 = \alpha_1^*$ *and* $\alpha_2 = \alpha_2^*$.

This proposition is valid for any sources configuration but the situation where $\mathbf{M}^*$, $\mathbf{X_{s_1}}$ and $\mathbf{X_{s_2}}$ are aligned. In this particular case where $\alpha_1^* = \alpha_2^*$ we get:

$${\mathbf{S}_2}^* = \begin{bmatrix} \cos\alpha_1 & \sin\alpha_1 & 0 \end{bmatrix}^\top. \tag{5.72}$$

As expected, we obtain a trajectory similar to the translation motion described for the case of a single source in (5.43), since only one angular data is available (see Figure 5.11b). For the stability, the condition $\mathbf{L}_\tau\widehat{\mathbf{L}_\tau^+} > 0$ is of course ensured when $\widehat{\ell_i} = \ell_i$ since we have $\mathbf{L}_\tau\widehat{\mathbf{L}_\tau^+} = \mathbb{I}_2$ in that case. Since ℓ_i is not available, a classical method to approximate each $\widehat{\ell_i}$ is to use the distance ℓ_i^* to the sound source at a desired pose. In this case, it is well known that the system is locally asymptotically stable in the neighborhood of the desired pose [CH08]. In our case, with an infinite set of desired poses, a simple choice is to fix $\widehat{\ell_1} = \widehat{\ell_2} = k$, where k is an approximation of the distance to the sources at the desired pose.

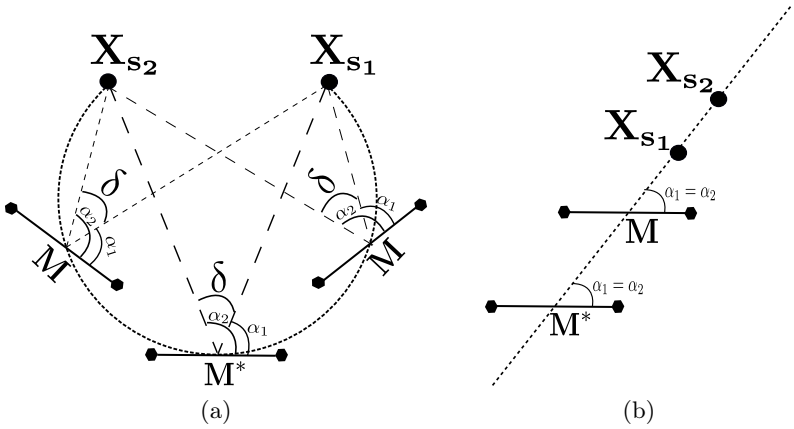


Figure 5.11: With two sound sources (a) several poses exist for given τ_1^* and τ_2^* on the circumscribed circle defined by the sound sources and the position of the microphones. When the sources are aligned with $\mathbf{M}$ (b), the circle becomes a line directed by the sources locations.

5.3.1.2 Experimental results

From the robot configuration given in Figure 5.5, the control scheme used in this experiment is based on the Jacobian feature matrix extracted from (5.70) as

$$\widehat{\mathbf{J}_{\boldsymbol{\tau}}} = \begin{bmatrix} \frac{\tau_1 \nu_1}{A\widehat{\ell_1}} & \frac{A\widehat{\ell_1}\nu_1 - D_x \nu_1^2}{A\widehat{\ell_1}} \\ \frac{\tau_2 \nu_2}{A\widehat{\ell_2}} & \frac{A\widehat{\ell_2}\nu_2 - D_x \nu_2^2}{A\widehat{\ell_2}} \end{bmatrix}. \tag{5.73}$$

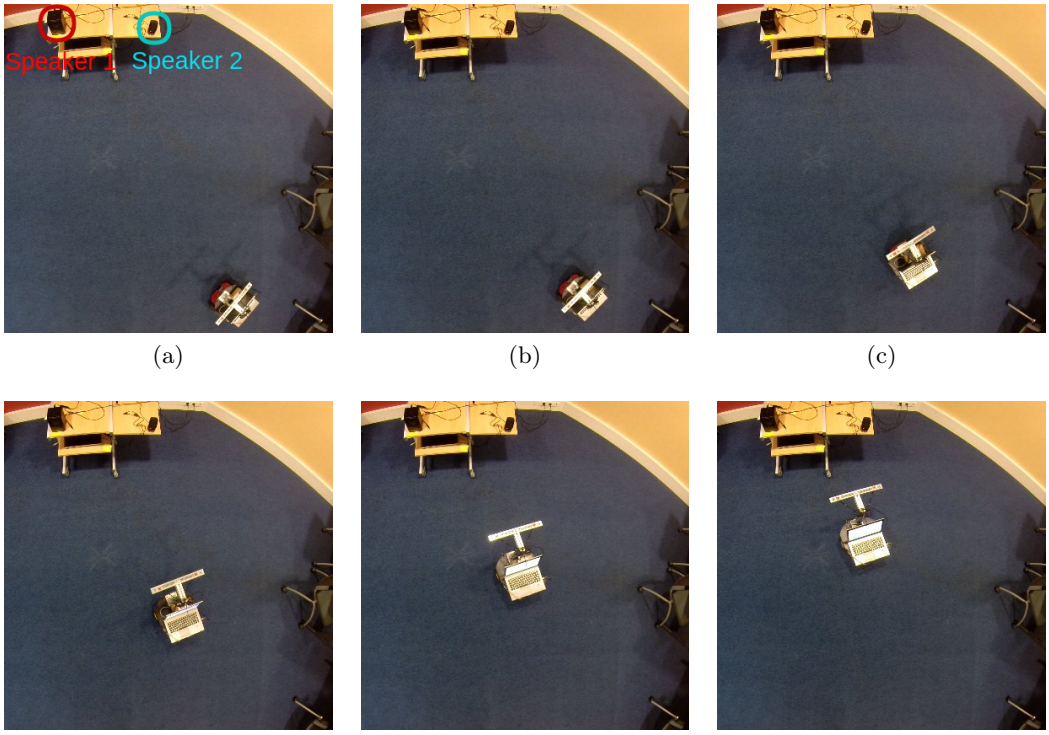
The control input $\dot{\mathbf{q}} = (u, \omega)$ is then computed from

$$\dot{\mathbf{q}} = -\lambda \widehat{\mathbf{J}_{\boldsymbol{\tau}}^{+}}(\boldsymbol{\tau} - \boldsymbol{\tau}^*), \tag{5.74}$$

in which $\boldsymbol{\tau} = [\tau_1, \tau_2]^\top$ and $\boldsymbol{\tau}^* = [\tau_1^*, \tau_2^*]^\top$. The settings are exactly the same as the case of a single source task and are given in Table 5.2. The approximated distances are $\widehat{\ell_1} = \widehat{\ell_2} = 1$ m. Furthermore the tracking routine described in Section 5.2.6.3 is used. Additionally to this tracking, a labelling of the retrieved ITDs is necessary when there are two sound sources. The goal is to associate each τ_i to the desired τ_i^* so that the task can be correctly completed. The labelling problem is trivial to solve in this configuration. If we consider the working space as the half plane in front of the microphones, the ordinality of $\tau_i(t)$ and τ_i^* is the same. Namely if $\tau_1^* < \tau_2^*$ then $\tau_1(t)$ should be lesser than $\tau_2(t)$. As shown in Figure 5.11a, it is obvious that each pose of the microphones in the environment can be characterized by a circumscribed circle on which the corresponding angles α_1 and α_2 have always the same ordinality.

Hence, in this experiment, besides the female speech we added a second sound source corresponding to a burst of white Gaussian noise of 25 ms followed by 25 ms of silence played in loop. This time, the objective is to reach a pose where $\tau_1^* = -\tau_2^*$ with $\alpha_1 = 50°$. From a pose fulfilling that condition, the system extracted τ_1^* and τ_2^*

in the first step. After, starting from a pose around 3 meters away from the sources, the system automatically initialized and labelled $\tau_1(t_0)$ and $\tau_2(t_0)$.



(a) (b) (c)

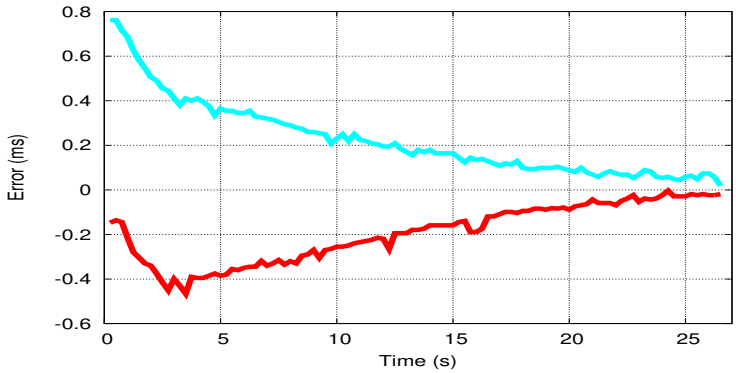
(d) (e) (f)

Figure 5.12: With two sound source, the robot heads towards a pose satisfying the given bearing conditions.

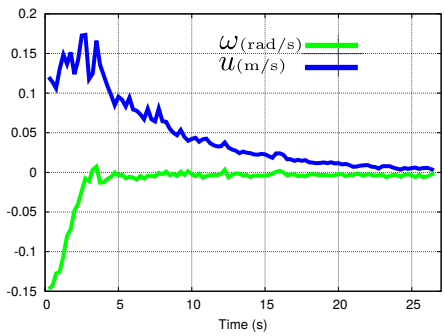
The Figure 5.12 shows the completion of the task. The data of this experiment (see Figure 5.13) shows corrupted ITD estimations occurring during the robot motion which were coped with the tracking routine. More precisely the spurious ITDs were caused by echoes following the same dynamics as the actual ITDs. This could be expected since the echoes appear as virtual sound sources. Moreover the noise effect is still present with observation of peaks for $\alpha = 90°$ during most of the frames. Despite these poor observations, the error of the measured ITDs successfully converges to zero while the robot followed a straight and smooth trajectory towards the circle set of solutions to be reached, with a correct orientation.

However, compared to the ITD-based positioning task using a single source, this type of positioning task is less robust. The robustness decreases notably because of the issue of tracking the ITDs related to the actual sources. When considering intermittent sound sources, the tracking algorithm does not account for the number of active sources at the current frame. Hence, some plausible ITDs issued from the echoes of the active source could be associated to the ITD of the inactive sound source. For this kind of configuration, it is then necessary to have a more reliable tracking that takes into consideration the number of active sound sources. Otherwise,

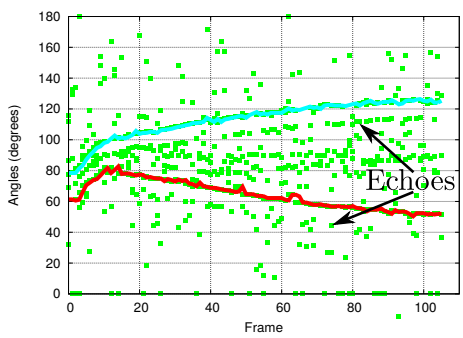
when considering continuous sound sources, the performance of the positioning task is as robust as the task considering a single sound source.



(a) Error $\boldsymbol{\tau} - \boldsymbol{\tau}^*$



(b) Control input



(c) Observed (dots) and tracked (lines) α_i deducted from ITDs

Figure 5.13: Experimental result of the ITD-based task with two sound sources illustrated in Figure 5.12.

5.3.2 Case of three sound sources and more

5.3.2.1 Control scheme and task analysis

It is also possible to consider more sound sources. When considering three sound sources $\mathbf{X_{s_1}}$, $\mathbf{X_{s_2}}$ and $\mathbf{X_{s_3}}$ the interaction matrix is generally full rank 3,

$$\mathbf{L}_{\boldsymbol{\alpha}} = \begin{bmatrix} \frac{\sin\alpha_1}{\ell_1} & -\frac{\cos\alpha_1}{\ell_1} & -1 \\ \frac{\sin\alpha_2}{\ell_2} & -\frac{\cos\alpha_2}{\ell_2} & -1 \\ \frac{\sin\alpha_3}{\ell_3} & -\frac{\cos\alpha_3}{\ell_3} & -1 \end{bmatrix} \tag{5.75}$$

For the ITDs τ_1, τ_2 and τ_3, related to each sound source, this interaction matrix becomes

$$\mathbf{L}_{\boldsymbol{\tau}} = \begin{bmatrix} -\frac{\nu_1^2}{A\ell_1} & \frac{\tau_1\nu_1}{A\ell_1} & \nu_1 \\ -\frac{\nu_2^2}{A\ell_2} & \frac{\tau_2\nu_2}{A\ell_2} & \nu_2 \\ -\frac{\nu_3^2}{A\ell_3} & \frac{\tau_3\nu_3}{A\ell_3} & \nu_3 \end{bmatrix} \tag{5.76}$$

In most cases, this configuration implies that there exists only one pose where $\mathbf{s} = \mathbf{s}^*$. However, depending on the specified desired position or the sound sources configuration, a set of poses for which $\mathbf{s} = \mathbf{s}^*$ might exist. Indeed, inspired by the case of two sound sources, a set can be expected if $\mathbf{X_{s_1}X_{s_2}X_{s_3}}$ and $\mathbf{M}$ design a concyclic quadrilateral. Wherever $\mathbf{M}$ belongs to the corresponding circumscribed circle, an orientation $\theta_{\mathbf{M}}$ exists such that $\mathbf{s} = \mathbf{s}^*$ (see Figure 5.14b) and $\mathbf{S}^*$ is designed by (5.71). This result can be demonstrated analytically by computing the determinant of the interaction matrix $\mathbf{L_\alpha}$:

$$|\mathbf{L_\alpha}| = \frac{\sin(\alpha_1 - \alpha_2)}{\ell_1\ell_2} + \frac{\sin(\alpha_2 - \alpha_3)}{\ell_2\ell_3} + \frac{\sin(\alpha_3 - \alpha_1)}{\ell_1\ell_3}. \tag{5.77}$$

Besides, from the inscribed angle theorem and the sinus law, it is possible to say that:

$$\frac{\|X_{s_3}X_{s_1}\|}{\sin(\alpha_3 - \alpha_1)} = \frac{\|X_{s_3}X_{s_2}\|}{\sin(\alpha_3 - \alpha_2)} = \frac{\|X_{s_2}X_{s_1}\|}{\sin(\alpha_2 - \alpha_1)}. \tag{5.78}$$

From now on, the determinant of the interaction matrix is equal to:

$$|\mathbf{L_\alpha}| = \frac{\sin(\alpha_3 - \alpha_1)(\ell_2 - \frac{\|X_{s_2}X_{s_1}\|}{\|X_{s_3}X_{s_1}\|}\ell_3 - \frac{\|X_{s_3}X_{s_2}\|}{\|X_{s_3}X_{s_1}\|}\ell_1)}{\ell_1\ell_2\ell_3}. \tag{5.79}$$

Furthermore, an other property of the cyclic quadrilateral is that the product of the diagonals equals to the sum of the products of opposite sides:

$$\ell_2\|X_{s_3}X_{s_1}\| = \ell_1\|X_{s_3}X_{s_2}\| + \ell_3\|X_{s_2}X_{s_1}\| \tag{5.80}$$

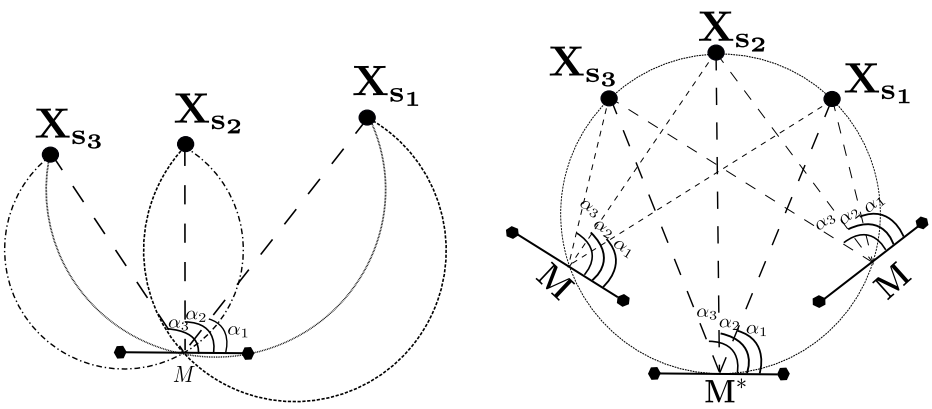
By combining equation (5.79) and (5.80), it appears that $|\mathbf{L_\alpha}| = 0$. As a result, there is a rank loss of the interaction matrix, and this system is similar to the case of two sound sources. There exists other configurations also leading to a rank loss of the matrix. By analyzing (5.77), $|\mathbf{L_\alpha}|$ is null when:

- $\alpha_1 = \alpha_2 = \alpha_3$. In that configuration, the sound sources are aligned in the microphones axis as $\overrightarrow{\mathbf{MX_{s_1}}} = k_1\overrightarrow{\mathbf{MX_{s_2}}} = k_2\overrightarrow{\mathbf{MX_{s_3}}}$, which corresponds to the configuration described by (5.72). But we can also notice that an infinitesimal motion defined by v_x, v_y or ω_z different from sound direction will cancel this singularity.

- $\alpha_1 = \alpha_2$, then we have $|\mathbf{L_\alpha}| = \frac{\sin\alpha_{12}\cos\alpha_3 - \cos\alpha_{12}\sin\alpha_3}{\ell_2\ell_3} + \frac{\cos\alpha_{12}\sin\alpha_3 - \sin\alpha_{12}\cos\alpha_3}{\ell_1\ell_3}$. As soon as $\ell_1 = \ell_2$, the determinant is equal to zero. Concretely, if at least two out of the three sources are juxtaposed, we are in the similar case of only two sound sources.

Apart from the aforementioned configurations, there is an unique solution for a task based on three sound sources (see Fig.5.14a). Geometrically, this result is obtained by considering the three circumscribed arcs of circle defined by $\mathbf{X_{s_1}X_{s_2}M}$, $\mathbf{X_{s_2}X_{s_3}M}$ and $\mathbf{X_{s_1}X_{s_3}M}$. According to Proposition 2, $\mathbf{M}$ must belong to these three arcs to ensure $\alpha_1 = \alpha_1^*$, $\alpha_2 = \alpha_2^*$ and $\alpha_3 = \alpha_3^*$. As shown in Figure 5.14b, these arcs intersect in only one position when considering a nominal configuration. From this result it can be set the following proposition.

Proposition 3 *In presence of three sound sources in a non-concyclic or non-degenerate configuration with* $\mathbf{M}^*$*, there exists only one pose* $\mathbf{M}$ *of the microphones so that* $\alpha_1 = \alpha_1^*$*,* $\alpha_2 = \alpha_2^*$ *and* $\alpha_3 = \alpha_3^*$*.*



(a) (b)

Figure 5.14: With three sound sources (a) a unique pose generally satisfies the conditions given by τ_1^*, τ_2^* and τ_3^*. When the sources and the microphones position $\mathbf{M}$ design a concyclic quadrilateral (b), there is set of solution equivalent to the configuration with two sound sources

Eventually, the circumscribed circle built by the sound sources is then a zone where the control scheme is singular and thus not stable. Indeed independently of the approximation $\widehat{\mathbf{L}_{\boldsymbol{\tau}}}$, the condition $\mathbf{L}_{\boldsymbol{\tau}}\widehat{\mathbf{L}_{\boldsymbol{\tau}}^{-1}} > 0$ cannot be ensured since at this position $|\mathbf{L}_{\boldsymbol{\tau}}| = 0$. We denote this zone as a singularity circle. A parallel can be emphasized with IBVS region of instability, where the circumscribed circle is the planar equivalent of the singular cylinder [MR+93].

5.3.2.2 Using more than three sound sources

Using more than three sound sources can be beneficial by giving more robustness to the control law. The demonstration developed for three sound sources remains applicable and a single pose exists such that $\mathbf{s} = \mathbf{s}^*$. Redundant data may ensure the stability of the control scheme when crossing the singularity circle that does nnot exist anymore. Indeed the singularity of the control scheme is cancelled by a fourth sound source positioned in a non-concyclic configuration with the three first. On the other hand, only the local asymptotic stability is verified under the condition $\widehat{\mathbf{L}_{\boldsymbol{\tau}}^{+}}\mathbf{L}_{\boldsymbol{\tau}} > 0$ (see Chapter 3.2.3), which is ensured when $\widehat{\ell_i} = \ell_i$ since $\widehat{\mathbf{L}_{\boldsymbol{\tau}}^{+}}\mathbf{L}_{\boldsymbol{\tau}} = \mathbb{I}_3$. Thence with three sound sources in a non-singular configuration or with at least four sources in any other case, all the three robot DOF are constrained. A homing task can be performed without an exact knowledge of the position of the sound sources (an approximation $\widehat{\ell_i}$ is sufficient). The realization of such task corresponds to bearing-only homing system.

Nonetheless, from an acoustic point of view, achieving such task remains extremely challenging when considering wideband and intermittent signal such as speech. In indoor environment, reverberation, sources (in)activity complicate the tracking and the identification of each ITD. Moreover, only few techniques are able to detect robustly such number of sound sources. The survey conducted in [BOV12] shows that GCC-PHAT has, at best, a recall rate of 0.7 for 3 sound sources. From this result, it can be inferred that it is likely that all the sources will not be detected in the sound mixture. For all these reasons, we think that such method should be particularly suitable for narrowband acoustic landmarks, that simplify the aforementioned problems. Each landmark is then detected on a limited and disjoint frequency range, which eases the detection and identification of the ITD. For instance under this hypothesis, in [BSLF16] the authors propose a follower-leader control scheme based on bearing measurements recorded by 4 microphones. In this configuration, the ITD-based positioning task could be fully exploited for multi-robot systems navigation or formation control.

5.3.2.3 Numerical evaluation

The developed method was tested through simulations in Roomsimove [BOV12]. We designed a room of 4.55 ×3.55 ×2.5 m^3 in which three sound sources $\mathbf{X_{s_1}}$, $\mathbf{X_{s_2}}$ and $\mathbf{X_{s_3}}$ were positioned as they appear in Figure 5.15a. Here, we simulated a holonomic robot that is controlled by the 3 DOF of the azimuth plane v_x, v_y and ω_z. First, each of the sound source emitted in succession so that it can be identified uniquely. Thereafter, the ITDs τ_i are estimated from the sound mixtures generated by the three sound sources. In the two first simulations described above we considered the real interaction matrix so that $\widehat{\mathbf{L}_{\boldsymbol{\tau}}^{-1}} = \mathbf{L}_{\boldsymbol{\tau}}^{-1}$ assuming that each distance ℓ_i is known. The effect of approximating the interaction matrix is discussed afterwards. The control scheme of the task is given by

$$\mathbf{u_M} = -\lambda \mathbf{L}_{\boldsymbol{\tau}}^{-1}(\boldsymbol{\tau} - \boldsymbol{\tau}^*), \tag{5.81}$$

where $\mathbf{L}_{\boldsymbol{\tau}}$ is defined by (5.76).

We simulated first a homing task with an initial pose defined by $\mathbf{M}(2, 0.75)$ and $\theta_{\mathbf{M}} = -5°$ and a home pose defined by $\mathbf{M}^*(3.1, 1.75)$ and $\theta_{\mathbf{M}^*} = 10°$. That home pose corresponds to $\boldsymbol{\alpha}^*$=$(31°, 85°, 118°)$ and was located outside of the circumscribed circle. In addition, a moderate reverberation was added with a reverberation rate $\mathrm{RT}_{60} \approx 75$ ms) and with background white noise that guarantees an SNR of 25 dB. The result of this simulation is shown on Figure 5.15. Despite the uncertainties on the ITDs measurement caused by the noise and reverberation, the system converged successfully to the home pose with an exponential decrease of the error. Moreover, since the singularity circle was not crossed during the task, the stability of the control law was ensured. As for simulation data, despite the jittery behaviour in the velocities curves caused by ITDS estimation, the trajectory remains smooth.

A second scenario was conducted including 99 trials on different steps of reverberation with uniformly distributed starting pose. With the *front-back ambiguity* implied by the use of only two microphones, $\theta_{\mathbf{M}}$ was chosen so that the sound sources

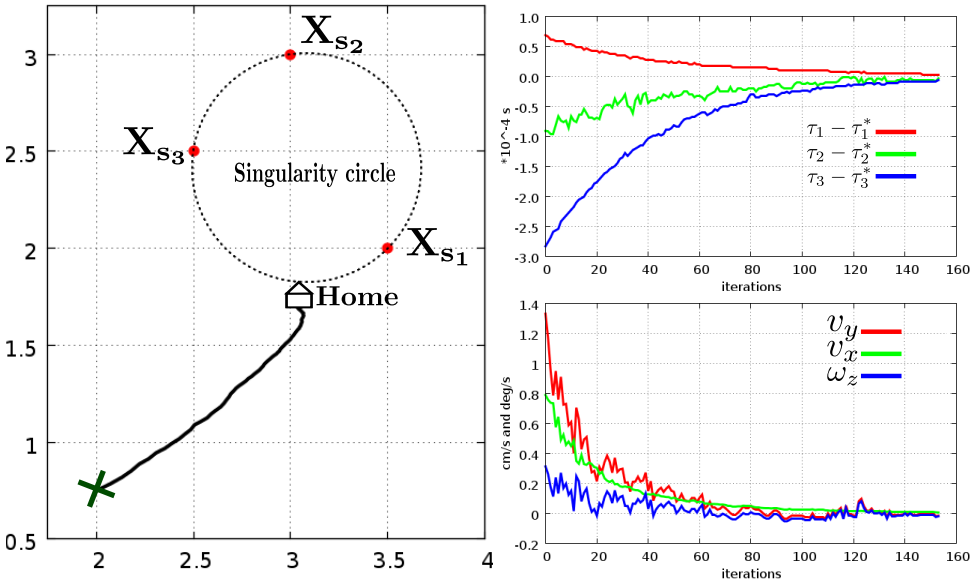
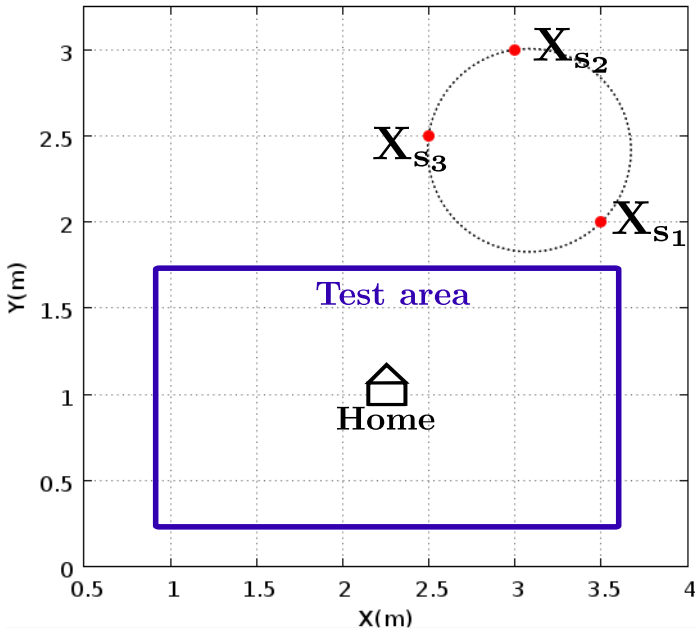


Figure 5.15: A typical homing task with moderate reverberation $\mathrm{RT}_{60} \approx 75$ ms

were on the same side of the microphones at the initial pose. Thus the orientation $\theta_{\mathbf{M}}$ varied from $-30°$ to $30°$ depending on the starting pose. With a home pose at $\mathbf{M}(2.25, 1)$ and $\theta_{\mathbf{M}^*} = -5°$, the results summed up in Figure 5.16 show that with no reverberation the system always converge to that pose. This result confirms the relevance and the suitability of the auditory modelling. However, we distinguished two failure areas in presence of reverberation. The first one was located on the alignment of $\mathbf{X_{s_2}}$ and $\mathbf{X_{s_3}}$ and mainly occurs when the reverberation increases. In this area, the main difficulty is to identify each measurement. With variation in the estimation and outlier (sound echoes) values, a wrong ITD could be associated to a sound source. Likewise, a feature loss was observed on the right extremity of the test area since one of the ITDs was not correctly estimated. Consequently the system could fail to converge in those two areas especially with high reverberation.



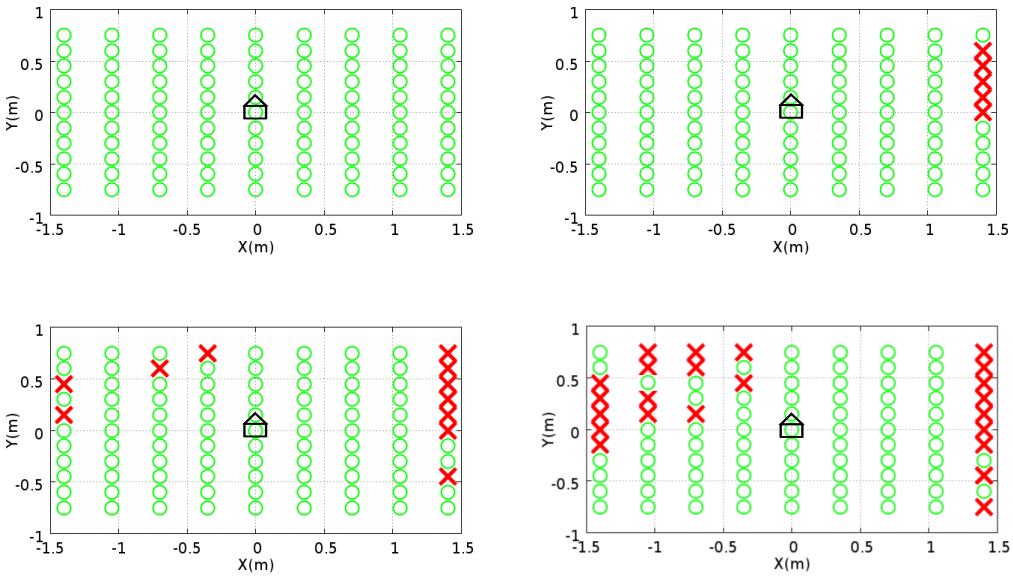


Figure 5.16: Homing task success in function of reverberation and initial pose: success(**O**) and failure (**X**) with from top-left to bottom-right with reverberation rate $\mathrm{RT}_{60} = \{0, 0.02, 0.05, 0.075\}$.

5.3.2.4 Approximation of the interaction matrix

Several ways exist to compute $\mathbf{L}_{\boldsymbol{\tau}}$ and these different strategies affect the behaviour and the stability of the system as shown in [MMR10]. Since the distances ℓ_i are unknown in practice, a first strategy consists in using ℓ_i^*, the distances at the desired pose, instead. As illustrated by Figure 5.17, $\mathbf{L}_{\boldsymbol{\tau}(\boldsymbol{\tau},\boldsymbol{\ell}^*)}$ gives a straighter and shorter trajectory. In addition, the desired measurements τ^* can be used in the interaction matrix. In that case $\mathbf{L}_{\boldsymbol{\tau}(\boldsymbol{\tau}^*,\boldsymbol{\ell}^*)}$ leads to a longer and unexpected trajectory but still converges to the home pose. These two solutions solve the problem of the unknown distance to the sound source. Moreover $\mathbf{L}_{\boldsymbol{\tau}(\boldsymbol{\tau}^*,\boldsymbol{\ell}^*)}$ is constant and needs to be computed only once. A last strategy considers the combination of the real and the constant interaction matrix leading to $\widehat{\mathbf{L}_{\boldsymbol{\tau}}} = (\mathbf{L}_{\boldsymbol{\tau}(\boldsymbol{\tau},\boldsymbol{\ell})} + \mathbf{L}_{\boldsymbol{\tau}(\boldsymbol{\tau}^*,\boldsymbol{\ell}^*)})/2$. The corresponding trajectory is improved compared to the constant interaction matrix, with the influence of $\mathbf{L}_{\boldsymbol{\tau}(\boldsymbol{\tau},\boldsymbol{\ell})}$.

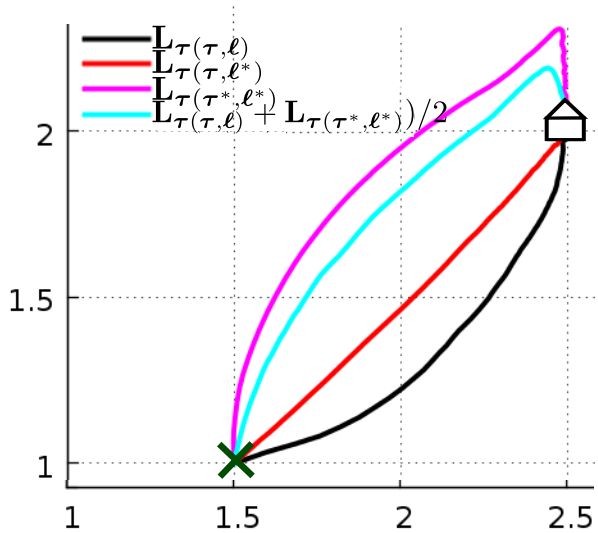


Figure 5.17: Homing trajectories with different interaction matrix approximations

5.4 Conclusion

In this chapter, we proposed a control scheme modelling based on either a single ITD or several ITDs in the configuration of multiple sound sources. These models are relying on the geometrical representation of ITDs. A given ITD expresses the sound source location through an hyperbola function. This hyperbola leads to two type of ambiguity that are the *left-right ambiguity* and the *front-back ambiguity*. The first ambiguity is solved from the sign of the measured ITD, but the *front-back ambiguity* remains unsolved, similarly to most binaural sound localization paradigms. Under far-field conditions, the ITD is directly related to the sound source direction, which is a common assumption in robot audition. This assumption corresponds geometrically to the branch of the hyperbola reaching their asymptotic values. However as the

sound source is getting closer to the microphones, namely in the near-field, this assumption does not hold anymore, since the hyperbola and its asymptotes are not merged. This result explains that most of the localization techniques based on ITD, only consider far-field configuration for accuracy purpose.

In a second hand, we focused more particularly on positioning tasks based on a single sound source. We then defined two interaction matrices that link the dynamics of a given ITD to the motion of the microphones. From virtual linkages approach, the positioning task was specified as orienting the robot with respect to the location of the source. Controller based on these matrices were then designed while respecting the stability conditions related to the approximations used. Similar conditions of stability were extracted from both ITD interaction matrices. These conditions are related to the *left-right ambiguity* (solved by the measurement of the ITD), and the *front-back ambiguity*. As expected from Chapter 3.1.1, the localization of the sound source is not required. The stability conditions are not constraining for both of these controllers. There exists an infinite number of solutions and the stability can easily be ensured by setting a positive source distance in the first matrix or by ensuring that the ambiguities expressed above are solved, in the case of the second matrix. Thus, by ensuring these conditions, experimental results validated the robustness of our approach. First, we demonstrated by analyzing the two interaction matrices, that our approach is not influenced by the distance of observation of the sound source. More explicitly, the positioning task can be performed in the near-field or far-field without restriction, despite rough approximations of the interaction matrices. This result is a clear advantage in comparison to binaural localization techniques that are generally devoted to only one type of configuration. Furthermore, experiments conducted on a mobile robot in a real environment emphasized the robustness of the controller. The robot is able to face and track an intermittent sound source that is moving. For this purpose, a basic tracking process was introduced. The tracking exploits the knowledge of the interaction matrix in order to predict the evolution of the tracked ITD. This is the principle of predictive control. A more exhaustive evaluation also confirmed that, thanks to this tracking, the control framework could be robust to real-world conditions since the controller can cope with false and missing measurements. Punctual errors and approximations in ITD tracking do not compromise the progress of the positioning task.

In the last segment, we extended the later control scheme to the case of multiple sound sources, that are used as acoustic landmarks for robot navigation. When considering several sound sources our approach is related to bearing-only navigation. Several positioning tasks can be achieved controlling up to three DOF in the azimuth plane. In a real-world experiment with two sound sources, we have been able to correctly position the robot with respect to the sound sources. The robot successfully reached a pose that belongs to a circumscribed circle shaped by the sound sources. This circle characterizes the desired bearing conditions. A single and stable pose of the microphones can be ensured with at least three sound sources as validated by simulations results. However, multi-source configurations require dedicated tracking algorithm in order to cope with adverse situations where competing and intermittent sound sources are present. This tracking is certainly the most critical aspect of

the control framework. The main issue concerns the tracking and identification of the ITDs associated to the actual source(s) rather than obtaining accurate intrinsic values.

Globally, the results exhibited in this chapter show robustness to inaccurate ITDs measurements. AS, as many other sensor-based control techniques, relies specifically on the dynamics of the auditory cues rather than the intrinsic value of the input feature. This property is confirmed in the next chapter, that demonstrates that the ILD framework and the current ITD framework could be freely applied on humanoid robots where the scattering effect of the head (*i.e.*, HRTFs) substantially modifies these cues.

Chapter 6

Application to humanoid robots

The last chapter of this thesis is concerned about the context of humanoid robots. Until now, AS paradigm has been applied in a free-field context. The free-field context greatly eases the auditory perception although real-world conditions are considered in this work. Nonetheless, all along the thread of this thesis several evidences showed the flexibility of AS. First the exact sound location is not needed to perform positioning tasks. The interaction matrices governing the motions of the robot are actually robust to rough approximated parameters while preserving stability properties. Furthermore, approximations of acoustic models are also well-tolerated. From a far-field model, positioning tasks in the near-field can be performed with the same accuracy as with a correct modelling. Eventually positioning tasks can be performed from cues (ILDs) that cannot be directly related to a precise source location.

Thus evidence converges towards the idea that accurate interaural cues (with respect to a ground truth source location) are not crucial to control the motion of the robots. This idea let us foresee a potential applicability of AS on humanoid robots. The main difficulty arising when considering humanoid robots comes from the required accurate modelling of HRTFs, and acoustic conditions, in order to relate the interaural cues to a sound location. However, since our approach is not attached to localize sound sources, thanks to the robustness of the control scheme, intuitively it can be envisaged that HRTFs modelling is not required for AS. This chapter aims to validate the latter idea. No new theoretical concepts are introduced here. We are rather interested in the application of the frameworks detailed in the previous chapters in the context of humanoid robots.

This chapter is then simply divided into two sections. In the first section, we briefly recall the context and the challenges of binaural hearing on humanoid robots. Following this review we give a comprehensive study on how the frameworks developed in Chapters 4 and 5 could be applied to humanoid robots. Thereafter, in the second section, experimental studies on two types of humanoid robots validate the versatility of AS paradigm. All these experiments are achieved considering a free-field modelling in real world environments.

6.1 Versatility of aural servo paradigm

Binaural localization on humanoid robots is the closest configuration to genuine auditory systems, but at the same time probably the most challenging configuration for robot audition. As explained in Chapter 2, in this configuration the combination of acoustic perturbations (reverberation, noise) and the scattering effect of the head creates challenging conditions to be resolved for sound localization. Very few works, in view of the size of robot audition community, tried to address sound localization in this context. So far, the approaches considered to address this configuration are threefold. The first set of solutions are based on the knowledge of HRTFs. Pre-measured HRTFs are used to infer the sound location. This kind of approaches is generally performed under anechoic/controlled conditions since the influence of room acoustic over HRTFs substantially modifies measurements. A second path consists in modelling the scattering effect of the head. For unknown robot HRTFs, the scattering theory, assuming a spherical rigid head is by far the most used model in robot audition. However this model does not account for acoustic conditions similarly to HRTF-based methods. Hence, these approaches suffer from the same limitations. The last type of approaches, based on learning methods, provides until now the best performances in realistic environment. But these methods are only locally efficient. The learning process cannot be generalized to every kind of robots or acoustic configurations. Learning all kinds of acoustic environments is not affordable due to the high variability and the dynamic nature of acoustic conditions. Currently, among the current projects related to robot audition only TWO!EARS (see Chapter 1.3) is addressing this challenging configuration. This project is based on the knowledge of *KEMAR* HRTFs given by the CIPIC database.

From AS perspective, in a first part, we focus on the question on how such approach could be adapted on humanoid robots. As stated in the scenario developed in Chapter 3, the core advantage of AS is certainly the fact that our approach is mainly based on the dynamics of the auditory features. As long as the dynamics of the selected features is consistent with the motion of the robot the desired task can be correctly achieved. This characteristic has been exemplified in Chapter 5 where we showed that the control scheme could be applied independently in the near-field or far-field despite approximated acoustic models. Similarly in Chapter 4 we demonstrated that a robot could face a sound source despite ILD estimations that could not be interpreted as an azimuth angle. Until now, in robot audition and more specifically in sound source localization, this kind of approach relying on the dynamics of the low-level features has not been fully exploited. As illustrated in the experiments performed all along this thesis, exploiting such characteristic increases the robustness to acoustic conditions. The dynamics of the auditory cues could also be robust to the individual variability of sound cues perception.

Unlike intrinsic values of the interaural cues that changes among each type of robots, the dynamics of the interaural cues is a consistent characteristic shared by all binaural setup. In order to assess the latter statement, we conducted a set of experiments based on ILD features. The experiments consist in observing the dynamics of a given ILD with respect to the source motions from different robots. Let

us first introduce the robots used in this experiment, that are respectively *Romeo* and *Pepper* from Softbank Robotics (see Figure 6.1). *Romeo* is a 1.4 m tall robot, equipped with four microphones and two pinnae. These microphones are embedded inside the robot head as represented in Figure 6.2. *Pepper* is a wheeled holonomic robot that is 1.2 m tall, also equipped with four microphones on the top of its head. However, unlike *Romeo*, this robot does not have pinnae. The HRTFs of both robots are unknown and are certainly quite different to each other due to their dissimilar structure and microphones topology. For both robots, we considered only two mi-

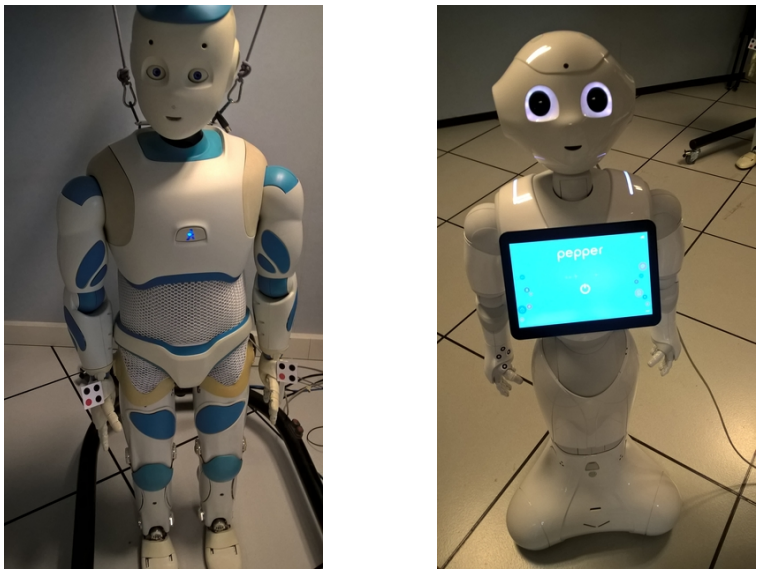


Figure 6.1: The two robots *Romeo* (right) and *Pepper* (left) used for measuring the dynamics of ILD cues

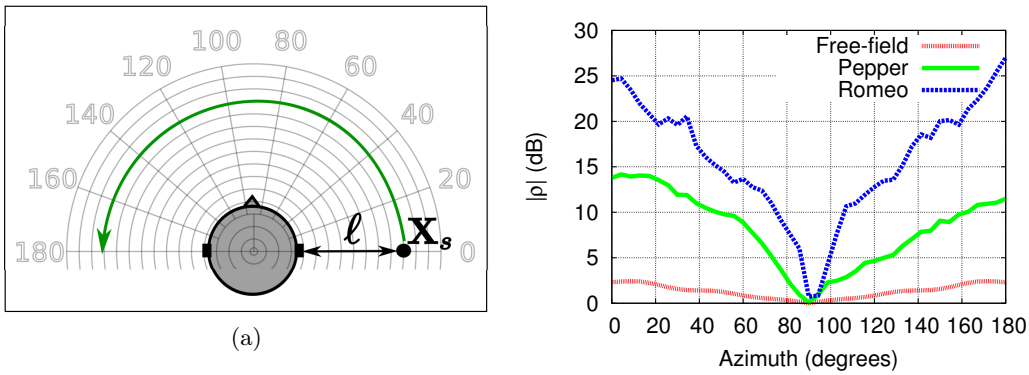
crophones $\mathbf{M_1}$ and $\mathbf{M_2}$ separated by a distance d, as depicted in Figure 6.2. The inter-microphones distances are respectively $d_{Pepper} \approx 0.07$ m and $d_{Romeo} \approx 0.12$ m. For the experiments, a white Gaussian noise is continuously emitting from a loudspeaker. This loudspeaker is moved in a circular motion around the robot head (*i.e.*, constant distance ℓ) so that the DOA α varies from 0° to 180°. The ILD ρ of the pair of microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ from the two robots is measured all along the motion of the loudspeaker.

These experiments are conducted in a highly reverberant environment with $\mathrm{RT}_{60} > 1$ s. Furthermore, the same task is conducted in simulation, however, in anechoic free-field conditions, with microphones separated by $d_f = 0.3$ m. The latter simulation serves as a baseline, for comparison with ideal conditions. The results are given in Figure 6.3 where the absolute value of $\rho_{dB} = 20\log_{10}(\rho)$ is plotted in order to facilitate the analysis.

First, it can be noticed that all configurations share the same "**V**" shape that reflects the symmetry of ILDs for sources at opposite positions with respect to the head. The ILD is maximal for eccentric positions of the source, while its minimum

Figure 6.2: Auditory system of *Romeo* and *Pepper*

value (≈ 0 dB) is reached when the source is in front of the robot. This property, shared by all configurations, is certainly the most interesting result for our approach. Despite different robot structures and acoustic conditions, the dynamics of the ILD is preserved. In contrast, the intrinsic value of ρ_{dB} is drastically changed depending on the auditory setup considered. Any approach based on localization would need to model or learn the influence of the HRTFs on the interaural cues. Hence, this kind of experiment confirms that independently of the robot platform, the dynamics of such acoustic feature remains consistent. The adverse experimental conditions ($RT_{60} > 1$ s), in the case of *Romeo* and *Pepper*, do not influence this property. The shape is the same for anechoic conditions and for the robots, despite a high level of reverberation. As a result, one can emphasize the robustness of our approach for real world conditions and more specifically to reverberation.



(a) (b)

Figure 6.3: $|\rho_{db}|$ measurements for a source at $\ell = 0.8$ m, for *Romeo*, *Pepper* obtained in real environment and for free-field microphones through simulation in anechoic conditions.

The details of the plotting in Figure 6.3 reveals some additional aspects of the sound perception. Among the three configurations tested, the "**V**" shape is more pronounced on *Romeo*. For this robot, the variation of the ILD is abrupt, while in free-field conditions this variation is more gradual. The shadowing effect of the head

increases the variation of ILDs as detailed in Chapter 2. The results of *Pepper* are in-between, but far more pronounced than for free-field conditions. The difference between *Romeo* and *Pepper* can be explained by the different microphones topology. The inter-microphone distance is wider for *Romeo* and the microphones are located on the lateral position of the head. The sound attenuation is then more pronounced compared to *Pepper* which endows microphones on the top of its head. From these results it can be inferred the following conclusion: head-mounted systems are better configurations than free-field systems for our approach. The shadowing effect of the head enhances ILD dynamics and any minimal variation of ILD (source motion) would be more "perceptible". As a consequence, the far-field limitations of ILD-based positioning tasks, described in Chapter 4, should be attenuated on humanoid robots. Furthermore, with enhanced ILDs, the effect of reverberation is reduced, especially for far-field conditions. This outcome is opposed with the commonly acknowledged idea in robot audition which stated that free-field configurations are easier to process.

Hence, from this simple evaluation, we deduce that the experiments performed in Chapter 4 can be performed on humanoid robots, with at least the same accuracy and without any additional modelling. Tasks consisting in facing a sound source could be performed on humanoid robots with a free-field modelling. It should also be noted that the same kind of results can be obtained when considering ITD cues. Therefore, the experiments depicted in Chapter 5 are also applicable on *Romeo* and *Pepper*. The experimental results in the following sections support these outcomes.

6.2 Experimental validation

6.2.1 Gaze control: facing a sound source

As a first case of study, this set of experiments consists in orienting the head of a robot towards a sound source. Despite the presence of the head between the microphones, the following experiments are based on free-field assumptions. More exactly we consider the modelling expressed in Chapter 4.1 on page 88 for the ILD case and in Chapter 5.1 on page 120 for the ITD case. The control inputs are given by

$$\dot{q}_{itd} = -\lambda \frac{1}{\sqrt{\left(\frac{d}{c}\right)^2 - \tau^2}} (\tau - \tau^*) \tag{6.1}$$

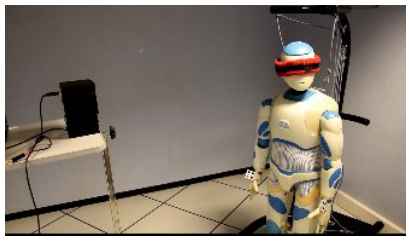
when controlling the robot through the ITD τ. From the ILD ρ the control input is given by

$$\dot{q}_{ild} = -\lambda \frac{\ell^2 + \frac{d^2}{4} - d\widehat{x}_s}{\widehat{y}_s d(\rho + 1)} (\rho - \rho^*). \tag{6.2}$$

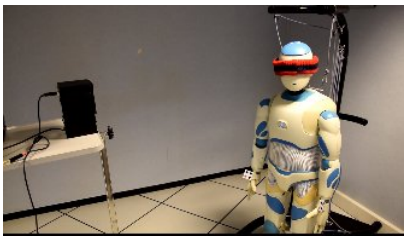
In these two control schemes, only the angular velocity ω, that sets the orientation of the robot head, is controlled.

The experimental setup consists in a speaker continuously emitting a white Gaussian noise. The acoustic of the room corresponds to the conditions described in the previous section. The room has a high reverberation rate $\mathrm{RT}_{60} > 1$ s and typical diffuse noises, related to the ventilation or computer fans, are present in the scene.

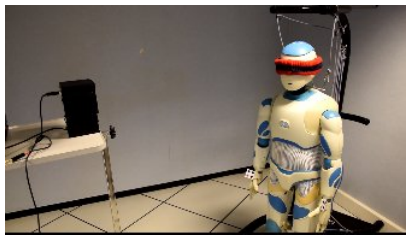
We consider beforehand the ILD case with the control input given by (6.2). For the task, we set $\rho^* = 1$. Furthermore, the approximated parameters appearing in the control scheme are set to $\widehat{x}_s = 1 \times sign(\rho - 1)$ and $\widehat{y}_s = 1$. The first experiment is carried on *Romeo*, with the use of the two microphones $\mathbf{M_1}$ and $\mathbf{M_2}$ as depicted in Figure 6.2. No knowledge of HRTFs nor modelling of the scattering effect of the robot, are considered in the following results presented in Figure 6.4. As predicted



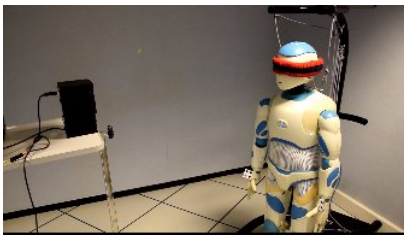
(a)



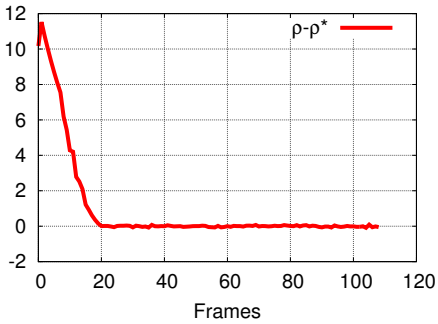
(b)



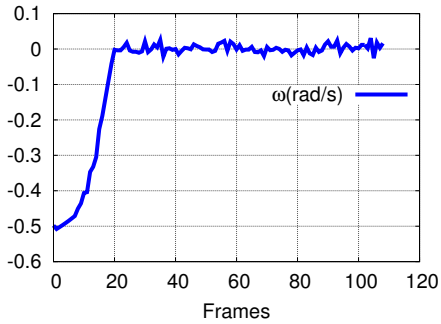
(c)



(d)



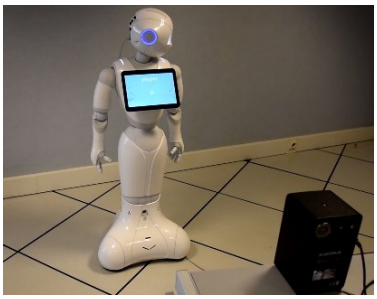
(e) Error



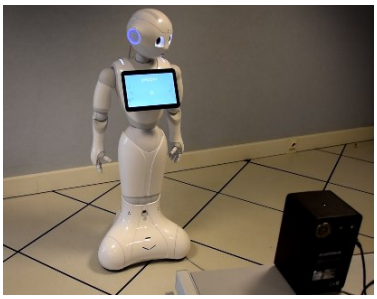
(f) velocity input

Figure 6.4: Head turning task from initial pose in (a) to final pose in (d) using ILD cues on *Romeo*.

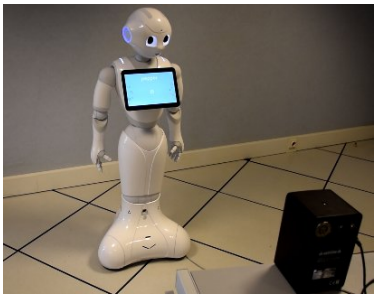
in Section 6.1, the head turning task is correctly achieved despite a modelling based on free-field conditions. The error curve follows an exponential decrease pattern while the robot accurately faces the sound source. This experiment has also been performed on Pepper in Figure 6.5. The same control scheme was used. Although the curves are jittery for both experiments when reaching the desired pose, the task is correctly achieved in both cases. Filtering strategies could also be applied on the measurements in order to smooth the motion of the robots. These experiments con-



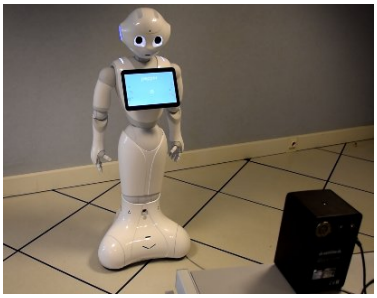
(a)



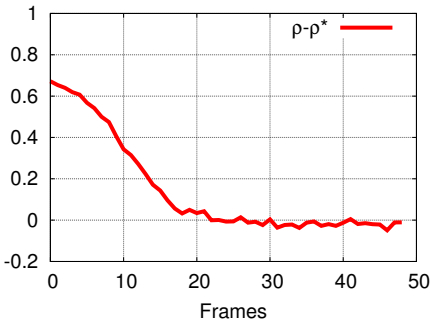
(b)



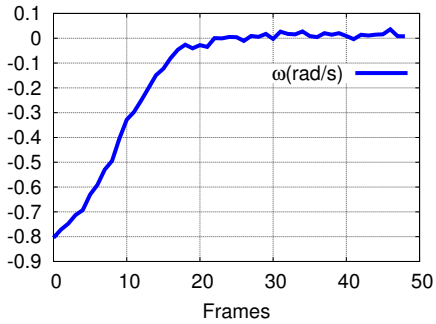
(c)



(d)



(e) Error



(f) velocity input

Figure 6.5: Head turning task from initial pose in (a) to final pose in (d) using ILD cues on *Pepper*.

firm the versatility of AS that can be applied to different humanoid robots without any additional modelling.

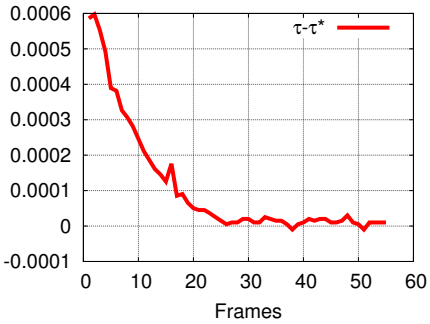
A second set of experiments were conducted, by using ITD measurements as input features for the positioning task. In this case, the control scheme for controlling the robot motion is given by (6.1). This experiment was performed on *Romeo* with a speaker emitting a continuous speech signal. Two external microphones were used instead of the embedded ones, because of the high level of internal noise in the robot. The GCC-PHAT algorithm is not particularly suitable for such level of noise. Despite the use of external microphones, the scattering effect of the head and the high level of reverberation in the scene are still influencing ITD measurements. Once again,

(a)

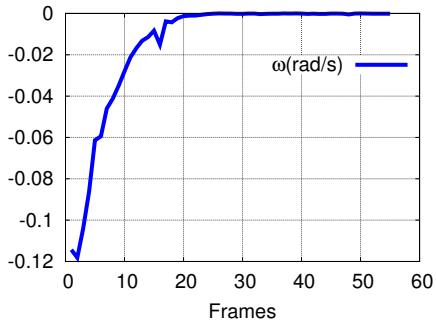
(b)

(c)

(d)



(e) Error



(f) velocity input

Figure 6.6: Head turning task from initial pose in (a) to final pose in (d) using ITD cues on *Romeo*.

the task is correctly achieved (see Figure 6.6) while using a non-stationary signal and without any knowledge of the scattering effect of the head. The control scheme is based on the AEG model where the azimuth angles are directly related to ITDs. For head-mounted systems, this model is obviously erroneous when considering localization approaches, however, it is correct when considering the dynamics of ITDs. Hence, AS greatly simplifies the auditory perception modelling, since AEG models can be used for various robotic platforms, for near-field and far-field distances, without any particular tuning.

6.2.2 Tracking a moving sound source

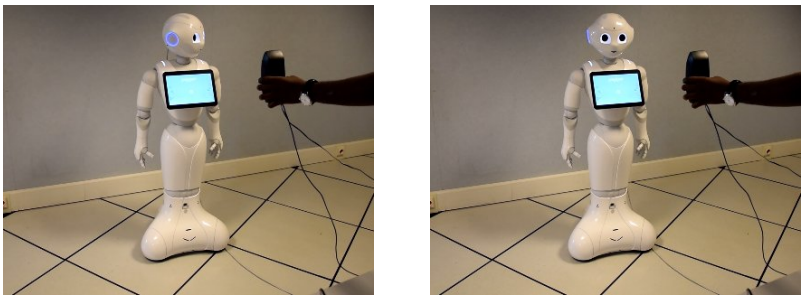
The control scheme is also able to track a moving sound source. In the experiment depicted in Figure 6.7, the speaker is continuously emitting a speech signal. The task consists in maintaining the sound source in front of *Romeo* head. ITD cues are used in this experiment with (6.1), but the same kind of results can be obtained with ILDs.

Figure 6.7: Tracking a continuously moving sound source with from (a) to final pose in (d) using ITD cues on *Romeo*.

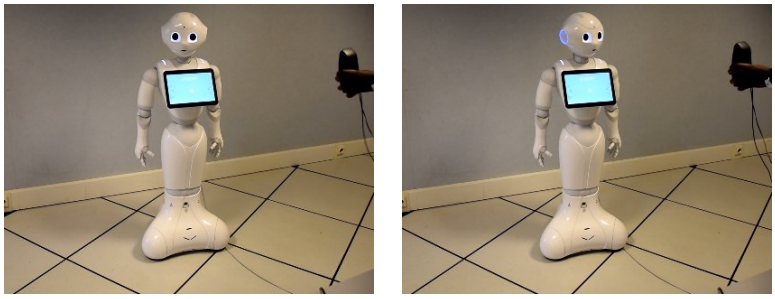
However, for ITD-based tasks, the speaker is moved more slowly than for ILD-based tasks. The tracking step of ITD limits the dynamics of the speaker: the track of the source could be lost with fast motions. On the other hand, since ILD-based tasks do not require any tracking, the robot is able to deal with faster motions of the sound source.

This property is confirmed by the second set of experiment achieved on *Pepper*. In this case the robot has to deal with rapid changes of the source position. From the controller, this experiment corresponds somehow to achieve a sequence of positioning tasks as depicted in Figure 6.8. Hence the error of the task follows a sequence of exponential decrease pattern.

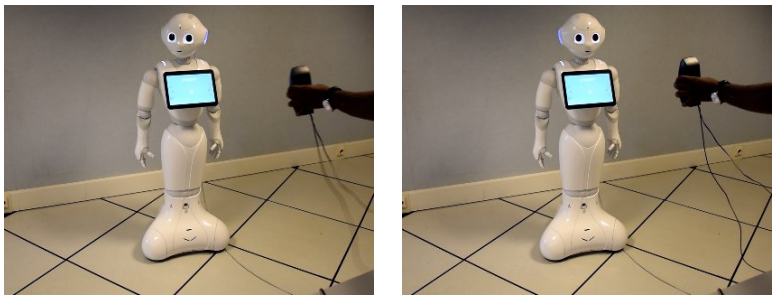
These two experiments were correctly achieved but we particularly noticed the effect of the sound source directivity. This effect is more visible when considering ILD-based control schemes. As already outlined in Chapter 4, the directivity of



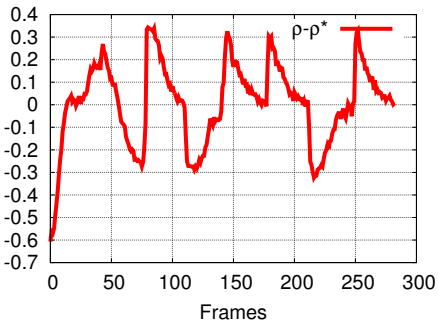
(a) (b)



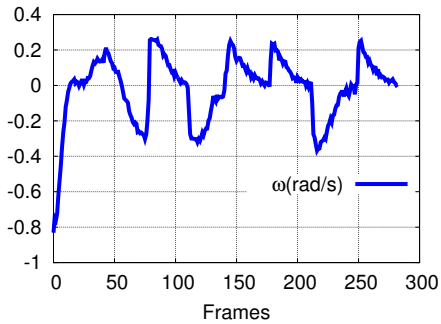
(c) (d)



(e) (f)



(g)



(h)

Figure 6.8: Experiment with rapidly changing source position using ILDs: is solved as a sequence of classic positioning task.

the sound source modifies the admissible poses for completing a given task. More precisely, since the sound wave is not emitted uniformly in all directions, the robot does not strictly aim to face the sound source. The task rather consists in facing the sound beam emitted by the speaker (see Figure 6.9). The controller correctly converges towards an appropriate configuration but perceptually the pose of the robot seems inaccurate.

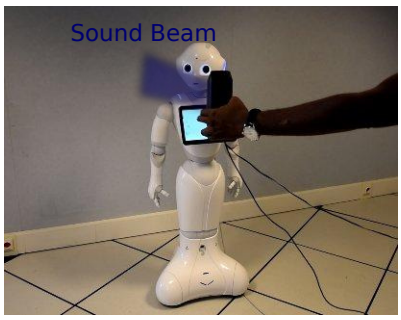


Figure 6.9: The effect of sound directivity: *Pepper* is facing the sound beam instead of the speaker

All along this manuscript, we mainly considered point sources uniformly emitting spherical waves, which is a commonly-admitted assumption in robot audition. In practice this assumption does not hold. All the sources emitting in realistic environment follow more complex radiation patterns governed by their intrinsic directivity property. A more thorough study is certainly required to include this notion in the control modelling.

6.2.3 Controlling Pepper with ILD and the energy level

For this last set of experiments, we aimed to apply the framework developed in Chapter 4.3, that considers ILDs and the energy level as input features. Here we are not attached to position the head of *Pepper*, but rather its holonomic base, that can be controlled with $\mathbf{u_M} = (v_x, v_y, \omega_z)$. In this case, the interaction matrix $\widehat{\mathbf{L}_{\rho E}}$ combining the ILD ρ to the energy level $E_\mathbf{M}$ can be directly used to control the robot. $\widehat{\mathbf{L}_{\rho E}}$ is given as follows:

$$\widehat{\mathbf{L}_{\rho E}} = \begin{bmatrix} \frac{2\widehat{x}_s(\rho-1)-d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{2\widehat{y}_s(\rho-1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} & \frac{\widehat{y}_s d(\rho+1)}{\widehat{\ell}^2+\frac{d^2}{4}-d\widehat{x}_s} \\ \frac{2E_\mathbf{M}\widehat{x}_s}{\widehat{\ell}^2} & \frac{2E_\mathbf{M}\widehat{y}_s}{\widehat{\ell}^2} & 0 \end{bmatrix}. \tag{6.3}$$

while the control scheme related to the velocity of the microphones is given by

$$\mathbf{u_M} = -\lambda \widehat{\mathbf{L}_{\rho E}}^{+}\mathbf{e}, \tag{6.4}$$

with $\mathbf{e} = [\rho - \rho^*, E_\mathbf{M} - E^*_\mathbf{M}]^\top$. In these experiments, the speaker is emitting a white Gaussian noise. The task consists in approaching the speaker from a random pose of the robot. For this purpose the desired ILD is set to $\rho^* = 1$. $E^*_\mathbf{M}$ is measured

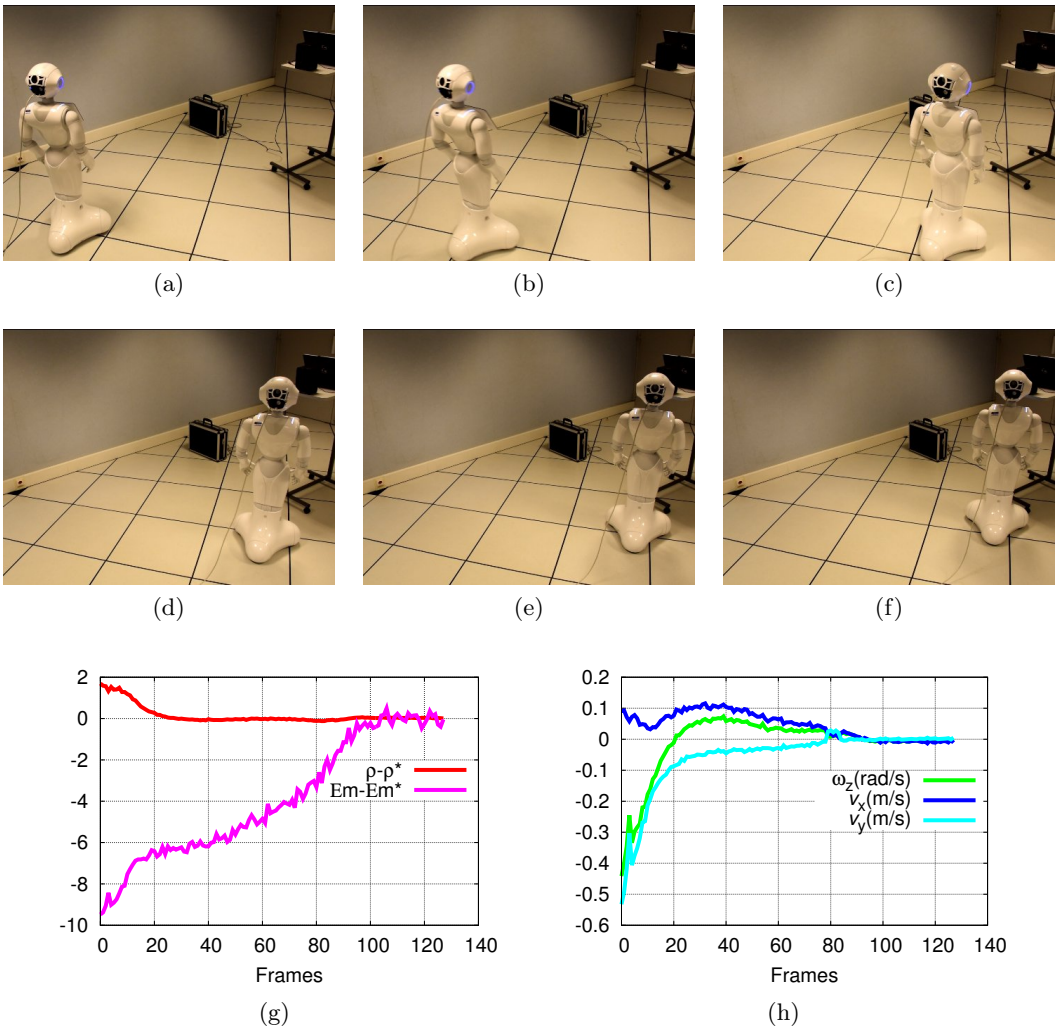


Figure 6.10: *Pepper* approaches the sound source and reaches a pose satisfying $\rho = \rho^*$ and $E_{\mathbf{M}} = E^*_{\mathbf{M}}$.

experimentally so that the robot is at $\ell \approx 0.3$ m from the sound source. The results are exposed in Figure 6.10. The robot is able to reach a pose satisfying the given auditory conditions. The behavior of the robot is, nonetheless, not really natural. The control scheme generates lateral motions that would be unsuitable for more realistic scenarios involving human-robot interaction for instance. Rotation motion would certainly give more naturalness to the approaching task. This behavior is due to the control that is uniquely applied to the base of the robot. Thus, it should be interesting to control at the same time the head (rotation) and the base of *Pepper* to get more natural and smooth motion. Despite these "unnatural" motions, from the latter application, the robot can also follow a moving sound source as illustrated in Figure 6.11.

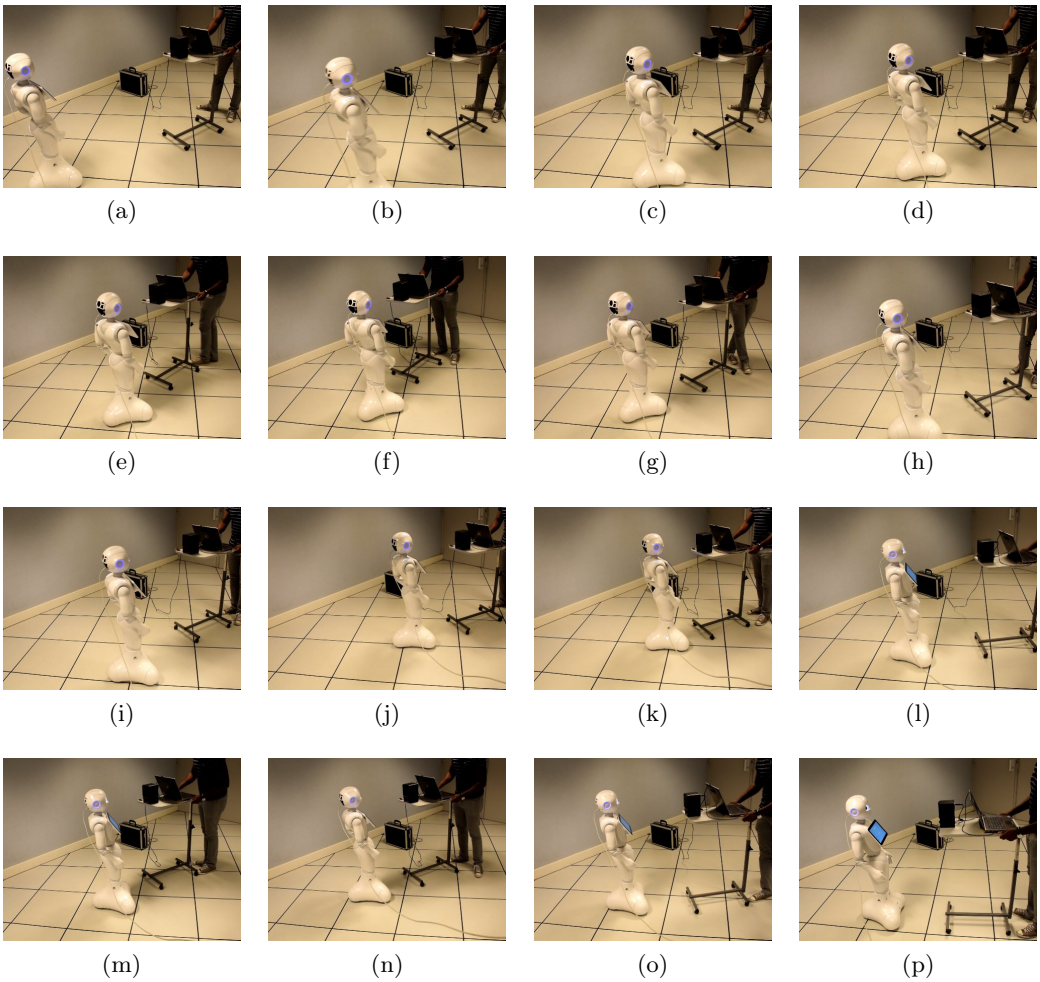


Figure 6.11: *Pepper* follows the sound source

6.3 Conclusion

This chapter confirmed the versatility of AS paradigm. This paradigm is mainly based on the dynamics of the selected auditory features. The dynamics of the auditory features is robust and flexible to changes of acoustic conditions and of binaural configurations. To demonstrate such characteristic, we showed, in a first phase, that the same dynamics governs ILD measurements for free-field anechoic conditions, and for head mounted systems in highly reverberant conditions. The same ILD shape could then be observed on two humanoid robots *Pepper*, *Romeo* and in simulated environments. From this result, it could be inferred that the framework developed in Chapters 4 and 5 considering free-field conditions can be applied on head mounted systems without any additional modelling.

Hence, in a second phase, we validated experimentally this framework on *Romeo* and *Pepper*. Despite the free-field modelling, convincing results could be obtained

on the humanoid robots, in realistic environments. A first set of experiments were concerned about controlling the gaze of the robots from ILD and ITD cues. The scattering effect of the head did not affect the accuracy of the task independently of the selected feature. Thereafter, the full motion control based on ILD and energy features was demonstrated to work accurately on *Pepper*. In this experiment, the translation and rotation velocities were controlled to approach a sound source. For all these experiments, the case of a moving sound source was also successfully addressed.

As the final chapter of this thesis, the experimental results depicted above sum up all the potential benefits of applying AS for real world applications. AS can be defined as a straightforward and simple paradigm that allows to control robots despite approximations in the auditory scene modelling and/or robot structure, and adverse acoustic conditions. Of course only basic scenarios and tasks have been addressed in this chapter, but these results unveil solid foundations for more complex and realistic scenarios.

Conclusion

Conclusions

As a recent field of research, robot audition encompasses all the topics related to the understanding and to the interaction with an acoustic scene. Sound source localization among others concentrated most of the efforts of the robot audition community. The localization paradigm initiates most of the applications based on the hearing sense. More specifically motion and behaviour control of the robot are mainly based on localization outputs. However, unlike many other fields related to perception (vision, touch/range sensing), applications in real-world context remains rare. The main bottleneck lies in controlling the robot from unreliable, inaccurate features and even sometimes from features that cannot be interpreted or related to an accurate sound location. These features are highly fluctuating and depends principally on the acoustic of the environment and the robot structure. Several parameters may change the perception of the sound. The reverberation, in indoor environment, and the noise substantially alter the auditory features for localization. Moreover, the distance of observation also impinges the auditory features. Acoustic waves heading towards the microphones are modelled differently in the far-field and the near-field. In the far-field it is easier to model the auditory features but on the other hand the effect of reverberation is more important. Alterations may also be generated by the structure of the robot. The scattering effect of the head, when considering anthropomorphic robots, and ego-noises are parameters to be included in the "equation" of sound localization. Eventually, when considering dynamic scenes, where the robot and/or the sound source of interest move, the degree of complexity of the localization task greatly increases. In view of all these parameters, sound localization in realistic and dynamic environments is a task that has not been achieved yet when considering binaural configurations as exposed in this thesis.

Hence, this thesis investigated the use of a sensor-based approach in order to control the motions of the robot with respect to auditory features. Sensor-based control corresponds to a feedback loop directly relating the motion of the robot to low-level auditory features. Unlike positioning methods based on the localization, here, the sensorimotor loop is closed and voluntary motions of the robot are driven by the auditory measurements. The resulting paradigm, that we termed *aural servo*, allows positioning the robot with respect to given conditions related to measured auditory features. More interestingly this approach is performed without localizing the source. This latter characteristic certainly represents a key advantage of *aural servo*.

Instead of focusing on localization that is extremely complex to solve in realistic environments, our approach focuses primarily on the dynamics of the auditory features. By discarding the localization concept to the positioning problem *aural servo* is more robust to parameters related to the environment or the robot morphology.

In this thesis, we introduced more specifically control models based on the ITD, the ILD and the absolute level of energy. From these models built upon the dynamics of these features with respect to the potential motion of the robot, various positioning tasks were performed. With both ILD and ITD, tasks that consists in orienting the robot and/or approaching a sound source are performed under realistic conditions. More impressively, the benefits of these frameworks have been emphasized by solving issues related to the auditory perception. In the case of the ILD, the front-back ambiguity is inherently solved by the control scheme. For ITDs, we demonstrated the robustness of the control scheme towards approximations in the distance of observation. And eventually, we proved that a range positioning using the absolute level of energy could be performed accurately under reverberant conditions. Furthermore these approaches are suitable for dynamic scenes. This problem is seldom addressed in robot audition. In this manuscript, from these approaches, we demonstrated the ability of our approach to control in real-time the orientation of the robot in order to face a moving sound source. It has also been demonstrated that the robot could accurately track a moving sound source under various changing acoustic conditions. More importantly, most of the concepts developed in this work are experimented on robots, in realistic acoustic environments. Besides stressing the suitability of *aural servo* for real environments, these experiments validate the versatility of the paradigm. The experiments can be carried, in the same way, on mobile robots equipped with free-field microphones or on humanoid robots where the microphones are integrated inside the head.

To the best of our knowledge, this work is one of the first attempts to characterize the motion of a robot with respect to the auditory information, in an explicit control framework. This manuscript is also supported by a large set of experiments in real world conditions, which is a core contribution of this thesis. Considering a binaural setup, the experimental results exhibited all along this thesis are not achievable yet by the state of the art of robot audition. Beyond this work, we would like to envision this approach as a general framework that can be applied to any type of robot or acoustic conditions. However, as a pioneering work, we are aware that this work only scratches the potential benefits of sensor-based control in robot audition. The methods presented in this thesis could be enhanced in order to interact, navigate, and understand the acoustic scene. In pursuance of this goal, this work should be complemented by the wide variety of methods developed for sound source localization, in order to handle intermittent sound sources, tracking and evaluating the number of active sources. Indeed, the path taken by this thesis is oriented towards the simplest and the most straightforward configurations that allows to control a robot from auditory feedback in real world conditions. As a result, this choice certainly emphasizes the actual benefits of using *aural servo* but on the other hand the solutions proposed in this manuscript might not be optimal and should be polished for more adverse and challenging scenarios.

Perspectives

Redefinition of the sound localization problem
More than an alternative approach to sound localization for robot motion control, we think that our approach presages a redefinition of sound localization problem and more generally interactions based on auditory perception. Until now, the common approach consists in making robust the interaural cues considering realistic environments and a particular auditory system. The localization and the tasks stemmed from it, solely depend on the accuracy of the extracted cues. This thesis showed that from inaccurate auditory cues, positioning tasks could be performed independently of the auditory system used. Hence, by adding *aural servo* in the mechanism of auditory interaction, the part devoted to localization is relaxed from the accuracy constraints, since *aural servo* can cope with inaccurate measurements. Such mechanism resembles to the human auditory system. As already explained in the introduction of this thesis, when using only interaural cues, humans are not so accurate in localizing sound compared to other species. However, by using the head motion, visual cues or the memory of previous auditory experiences, the human auditory system reveals all its potential and is then fully exploited.

Similarly, here, we could imagine a global framework where the auditory cues are extracted, tracked, identified and fed to an *aural servo* mechanism in order to interact with a given sound source. In this way, addressing the cocktail-party problem could be performed by combining *aural servo* with a source activity detection algorithm, methods for inferring the number of sound sources, robust tracking and an attention mechanism. These methods have already been studied and developed in the signal processing/robot audition literature in the context of sound localization. Most of these methods could be used in synergy with the ILD-based or ITD-based positioning tasks. Hence the "localization" problem would rather be a problem of detection and tracking.

An unified framework
In this work, ILDs and ITDs have been treated separately. As a logic extension, combining ILDs and ITDs in an unified framework can be envisaged. As explained in the first chapter, the human auditory system combines several auditory features in order to localize the potential sound source. In our context, although ILD-based tasks and ITD-based tasks are redundant with respect to the task that can be achieved, these tasks exhibit complementary characteristics. ILD-based methods are particularly adapted to single source tasks, and do not require any tracking. On the other hand, ITD-based methods can handle multi-source configurations and non stationary/intermittent signals such as speech. However, when using ITDs, the bottleneck concerns the identification and tracking of the sound source. Combining these two features could then overcome the limitations of both methods. Multi-modal control and hybrid sensor-based control are certainly paths to explore to achieve this task. We can then imagine using all the cues involved in sound localization, in a control scheme positioning the robot with respect to the azimuth, the elevation and the distance to the sound source.

Integration with other modalities

Aural servo could also be extended by using other sensory modalities such as vision. Interactions as performed by humans generally combine several senses. Hearing sense complements the limited field of view of vision sense, while vision is useful to identify brief and/or intermittent sound sources. The same complementarity could be exploited for robots. In this thesis, we considered that the source was long enough to allows converging to the desired pose. But in realistic context, it is not always the case. A thorough study should then be conducted in order to define the optimal time to convergence of such approach in comparison to the mean duration of a sound signal. In this case, predictive control or techniques such as visual servoing (when the source is visually identifiable) could be used if the controller has not converged at the end of the sound stimulus.

Extension to other auditory cues

One of the benefit of our approach lies in the fact that we do not need to localize the sound source. This interesting property opens perspective about using auditory cues that are for now neglected since they cannot easily be linked to a source location. For instance, the direct-to-reverberant ratio could be exploited in this manner to position the robot at a given range to the source. Spectral notches, related to HRTFs and the elevation of the sound source could be also exploited in the same manner, so that the control scheme is not limited to the azimuth plane. In the same vein, dense methods could be developed to directly interpret the sound signal and skip the stage of auditory cues estimation. This approach could be really useful in the case of ITDs by allowing to skip the tracking step. An explicit method consists in deriving the motion of the robot directly from the cross-correlation function, so that the tracking and identification step is avoided. Such approach has already been used with vision sense through mutual information [DM11] or photometric moments [BCM13].

During this thesis several paths and idea of robot audition applications have been explored. Among them, a navigation system based on echoes was an interesting idea that could not be fully developed, principally for of lack of time. In this configuration, the echoes are used as input features of a sensor-based control framework. From these echoes used as range sensing, the robot is able to position itself with respect to walls similarly to bats.

Extension of the field of applications

As underlined in the multi-sources simulations in Chapter 5, bearing-only navigation can be performed with sound and more particularly ITDs. Unfortunately, this method could not be fully experimented during the thesis. Applications of this method in the field of multi-robot control or autonomous navigation could be particularly interesting to expand the field of robot audition. We do believe that this kind of approaches could be implemented without a lot of effort. Moreover, these approaches allow connecting robot audition to fields with a wider exposure. Such connectivity could be beneficial for robot audition by increasing the awareness of the robotic community about hearing sense and the ensued problematic.

Furthermore *aural servo* is not limited to binaural configurations. Array-based configurations could also be considered in dedicated control framework. Such application could be a consistent alternative to usual localization techniques, depicted

in Chapter 2.3, that are principally based on the fusion of information. This fusion step does not appear in *aural servo* paradigm. Hence, better processing time should be obtained with similar flexibility and robustness as the tasks described in this manuscript.

However, on the path of extending our approach, more evaluations are required. The development of appropriate simulation tools is essential. Currently most acoustic simulated environments are designed for punctual and static situations. These tools are particularly slow when simulating high level of reverberation. As a result most of simulation conducted in this manuscript are based on moderate reverberation, since higher level of reverberation is not affordable in term of processing time. A real-time simulation tool is mandatory to process deeper analysis of the control frameworks.

Bibliography

[AAM99] F. Asono, H. Asoh, and T. Matsui. Sound source localization and signal separation for office robot "jijo-2". In *IEEE/SICE/RSJ International Conference on Multisensor Fusion and Integration for Intelligent Systems, 1999.*, pages 243–248. IEEE, 1999.

[AB79] J B. Allen and D A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[AC+] D. Abran-Côté et al. Usb synchronous multichannel audio acquisition system. Technical report.

[ADN95] D H. Ashmead, D L. Davis, and A. Northington. Contribution of listeners' approaching motion to auditory distance perception. *Journal of Experimental Psychology: Human Perception and Performance*, 21(2):239, 1995.

[ADS06] S. Argentieri, P. Danes, and P. Soueres. Modal analysis based beamforming for nearfield or farfield speaker localization in robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 866–871. IEEE, 2006.

[ADS15] S. Argentieri, P. Danes, and P. Soueres. A survey on sound source localization in robotics: From binaural to array processing methods. *Computer Speech & Language*, 34(1):87–112, 2015.

[ADTA01] V R. Algazi, R O. Duda, D M. Thompson, and C. Avendano. The cipic hrtf database. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 99–102. IEEE, 2001.

[AEB06] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[AGF+99] S Aharinejad, K Grossschmidt, P Franz, J Streicher, F Nourani, CA MacKay, W Firbas, H Plenk, and SC Marks. Auditory ossicle abnormalities and hearing loss in the toothless (osteopetrotic) mutation in the rat and their improvement after treatment with

colony-stimulating factor-1. *Journal of Bone and Mineral Research*, 14(3):415–423, 1999.

[Agi77] G J. Agin. *Servoing with visual feedback.* SRI International, 1977.

[Agi79] G J. Agin. *Real time control of a robot with a mobile camera.* SRI International, 1979.

[AHH+97] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui. Socially embedded learning of the o ce-conversant mobile robot jijo-2. In *International Joint Conference on Artificial Intelligence*, volume 97, pages 880–885. Citeseer, 1997.

[ALO90] D. H Ashmead, D. Leroy, and R D Odom. Perception of the relative distances of nearby sound sources. *Perception & Psychophysics*, 47(4):326–331, 1990.

[ANK+99] A Alford, S Northrup, K Kawamura, KW Chan, and J Barile. A music playing robot. In *Proceedings of the Conference on Field and Service Robots*, pages 29–31, 1999.

[APH12] X. Alameda-Pineda and R. Horaud. Geometrically-constrained robust time delay estimation using non-coplanar microphone arrays. In *Proceedings of the 20th European Signal Processing Conference*, pages 1309–1313. IEEE, 2012.

[APH14] Xavier Alameda-Pineda and Radu Horaud. A geometric approach to sound source localization from time-delay estimates. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(6):1082–1095, 2014.

[APSRH13] X. Alameda-Pineda, J. Sanchez-Riera, and R. Horaud. Benchmarking methods for audio-visual recognition using tiny training sets. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3662–3666. IEEE, 2013.

[AT00] F Augusztinovicz and M Tournour. Reconstruction of source strength distribution by inversing the boundary element method. *Boundary Elements in Acoustics, Advances and Applications*, 2000.

[AVS15] G. Athanasopoulos, W. Verhelst, and H. Sahli. Robust speaker localization for real-world robots. *Computer Speech & Language*, 34(1):129–153, 2015.

[B+16] J G. Betts et al. Anatomy & physiology. *Open Stax College*, 2016.

[BADM09] J. Bonnal, S. Argentieri, P. Danès, and J. Manhès. Speaker localization and speech extraction with the ear sensor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 670–675. IEEE, 2009.

[Bat67] D W. Batteau. The role of the pinna in human localization. *Proceedings of the Royal Society of London B: Biological Sciences*, 168(1011):158–180, 1967.

[BC94] G J Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336, 1994.

[BCM13] M. Bakthavatchalam, F. Chaumette, and E. Marchand. Photometric moments: New promising candidates for visual servoing. In *IEEE International Conference on Robotics and Automation*, pages 5241–5246. IEEE, 2013.

[BDFP16] G. Bustamante, P. Danes, T. Forgue, and A. Podlubne. Towards information-based feedback control for binaural active localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6325–6329. IEEE, 2016.

[BG05] S. Birchfield and R. Gangishetty. Acoustic localization by interaural level difference. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages iv–1109. IEEE, 2005.

[BHR09] J. Broekens, M. Heerink, and H. Rosendal. Assistive social robots in elderly care: a review. *Gerontechnology*, 8(2):94–103, 2009.

[BIK+15] Y. Bando, K. Itoyama, M. Konyo, S. Tadokoro, K. Nakadai, K. Yoshii, and H G Okuno. Microphone-accelerometer based 3d posture estimation for a hose-shaped rescue robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5580–5586. IEEE, 2015.

[Bla80] J Blauert. Modeling of interaural time and intensity difference discrimination. *Psychophysical, Physiological, and Behavioural Studies in Hearing*, pages 412–424, 1980.

[Bla97] J. Blauert. *Spatial hearing: the psychophysics of human sound localization.* MIT press, 1997.

[BMH03] J P. Barreto, F. Martin, and R. Horaud. Visual servoing/tracking using central catadioptric images. In *Experimental Robotics VIII*, pages 245–254. Springer, 2003.

[BNP+10] M. Bernard, S. N'Guyen, P. Pirim, B. Gas, and J-A Meyer. Phonotaxis behavior in the artificial rat psikharpax. In *International Symposium on Robotics and Intelligent Sensors*, pages 118–122, 2010.

[Bod93] M. Bodden. Modeling human sound-source localization and the cocktail-party-effect. *Acta Acoustica*, 1:43–55, 1993.

[BOV12] C. Blandin, A. Ozerov, and E. Vincent. Multi-source tdoa estimation in reverberant audio using angular spectra and clustering. *Signal Processing*, 92(8):1950–1960, 2012.

[BPDCG12] M. Bernard, P. Pirim, A. De Cheveigné, and B. Gas. Sensorimotor learning of sound localization from an auditory evoked behavior. In *IEEE International Conference on Robotics and Automation*, pages 91–96. IEEE, 2012.

[BR99] D S. Brungart and W M. Rabinowitz. Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479, 1999.

[Bre94] A S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[BS97] M. Brandstein and H F. Silverman. A practical methodology for speech source localization with microphone arrays. *Computer Speech & Language*, 11(2):91–126, 1997.

[BSFL14] M. Basiri, F. Schill, D. Floreano, and P. U Lima. Audio-based localization for swarms of micro air vehicles. In *IEEE International Conference on Robotics and Automation*, pages 4729–4734. IEEE, 2014.

[BSLF12] M. Basiri, F. Schill, P U. Lima, and D. Floreano. Robust acoustic source localization of emergency signals from micro air vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4737–4742. IEEE, 2012.

[BSLF16] M. Basiri, F. Schill, P. Lima, and D. Floreano. On-board relative bearing estimation for teams of drones using sound. *IEEE Robotics and Automation Letters*, 1(2):820–827, 2016.

[BT68] J C. Boudreau and C. Tsuchitani. Binaural interaction in the cat superior olive s segment. *Journal of Neurophysiology*, 31(3):442–454, 1968.

[BVMA09] A. Badali, J-M. Valin, F. Michaud, and P. Aarabi. Evaluating real-time audio localization algorithms for artificial audition in robotics. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2033–2038. IEEE, 2009.

[C+96] P. Corke et al. *Visual Control of Robots: high-performance visual servoing*. Research Studies Press Baldock, 1996.

[CB02] I. Cohen and B. Berdugo. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE signal processing letters*, 9(1):12–15, 2002.

[CB05] L. Chittka and A. Brockmann. Perception space—the final frontier. *PLoS Biol*, 3(4):e137, 2005.

[CBL+09] H Christensen, J Barker, YC Lu, J Xavier, R Caseiro, and H Arafajo. Popeye: Real-time, binaural sound source localization on an audio-visual robot-head. In *Conference on Natural Computing and Intelligent Robotics*, 2009.

[CC87] BA Cartwright and TS Collett. Landmark maps for honeybees. *Biological cybernetics*, 57(1-2):85–93, 1987.

[CCW06] W. Cui, Z. Cao, and J. Wei. Dual-microphone source location method in 2-d space. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages IV–IV. IEEE, 2006.

[CFC02] J H. Casseday, T. Fremouw, and E. Covey. The inferior colliculus: a hub for the central auditory system. In *Integrative functions in the mammalian auditory pathway*, pages 238–318. Springer, 2002.

[CG93] P Corke and MC Good. Controller design for high performance visual servoing. In *12th World congress IFAC*, volume 9, pages 395–398, 1993.

[CH06] F. Chaumette and S. Hutchinson. Visual servo control. i. basic approaches. *IEEE Robotics and Automation Magazine*, 13(4):82–90, 2006.

[CH08] F. Chaumette and S. Hutchinson. Visual servoing and visual tracking. In *Springer Handbook of Robotics*, pages 563–583. Springer, 2008.

[Che53] E C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[Che61] C. Cherry. Two ears–but one world. *Sensory communication*, pages 99–117, 1961.

[CK83] D Caird and R Klinke. Processing of binaural stimuli by cat superior olivary complex neurons. *Experimental Brain Research*, 52(3):385–399, 1983.

[CK88] CE Carr and M Konishi. Axonal delay lines for time measurement in the owl's brainstem. *Proceedings of the National Academy of Sciences*, 85(21):8311–8315, 1988.

[CLH97] S. Carlile, P. Leong, and S. Hyams. The nature and distribution of errors in sound localization by human listeners. *Hearing research*, 114(1):179–196, 1997.

[CLLH07] M. Cooke, Y-C Lu, Y. Lu, and R. Horaud. Active hearing, active speaking. In *ISAAR 2007-International Symposium on Auditory and Audiological Research*, pages 33–46, 2007.

[CMPC06] A. Comport, E. Marchand, M. Pressigout, and F. Chaumette. Real-time markerless tracking for augmented reality: the virtual visual servoing framework. *IEEE Transactions on visualization and computer graphics*, 12(4):615–628, 2006.

[CMS11] A. Comport, R. Mahony, and F. Spindler. A visual servoing model for generalised cameras: Case study of non-overlapping cameras. In *IEEE International Conference on Robotics and Automation*, pages 5683–5688. IEEE, 2011.

[CMWB09] H. Christensen, N. Ma, S. N Wrigley, and J. Barker. A speech fragment approach to localizing multiple speakers in reverberant environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4593–4596. IEEE, 2009.

[Col63] P. Coleman. An analysis of cues to auditory depth perception in free space. *Psychological Bulletin*, 60(3):302, 1963.

[Col73] H S. Colburn. Theory of binaural interaction based on auditory-nerve data. i. general strategy and preliminary results on interaural discrimination. *The Journal of the Acoustical Society of America*, 54(6):1458–1470, 1973.

[Col77] H S. Colburn. Theory of binaural interaction based on auditory-nerve data. ii. detection of tones in noise. *The Journal of the Acoustical Society of America*, 61(2):525–533, 1977.

[Cor03] P. Corke. Mobile robot navigation as a planar visual servoing problem. In *Robotics Research*, pages 361–372. Springer, 2003.

[CP94] S. Carlile and D. Pralong. The location-dependent nature of perceptually salient features of the human head-related transfer functions. *The Journal of the Acoustical Society of America*, 95(6):3445–3459, 1994.

[CRE93] F. Chaumette, P. Rives, and B. Espiau. Classification and realization of the different vision-based tasks. *Visual Servoing*, 7:199–228, 1993.

[CTS68] P. Cochran, J. Throop, and WE Simpson. Estimation of distance of a source of sound. *The American journal of psychology*, 81(2):198–206, 1968.

[DC77] RH Domnitz and HS Colburn. Lateral position and interaural discrimination. *The Journal of the Acoustical Society of America*, 61(6):1586–1598, 1977.

[DH97] CJ Darwin and RW Hukin. Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *The Journal of the Acoustical Society of America*, 102(4):2316–2324, 1997.

[DH12] A. Deleforge and R. Horaud. The cocktail party robot: Sound source separation and localization with an active binaural head. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 431–438. ACM, 2012.

[DHD07] K D. Donohue, J. Hannemann, and H G. Dietz. Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments. *Signal Processing*, 87(7):1677–1691, 2007.

[DHSG15] A. Deleforge, R. Horaud, Y. Schechner, and L. Girin. Co-localization of audio sources in images using binaural features and locally-linear regression. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(4):718–731, 2015.

[DK15] A. Deleforge and W. Kellermann. Phase-optimized k-svd for signal extraction from underdetermined multichannel sparse mixtures. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 355–359. IEEE, 2015.

[DLORG08] A. De Luca, G. Oriolo, and P. Robuffo Giordano. Feature depth observation for image-based visual servoing: Theory and experiments. *The International Journal of Robotics Research*, 27(10):1093–1116, 2008.

[DM98] R O. Duda and W L. Martens. Range dependence of the response of a spherical head model. *The Journal of the Acoustical Society of America*, 104(5):3048–3058, 1998.

[DM11] A. Dame and E. Marchand. Mutual information-based visual servoing. *IEEE Transactions on Robotics*, 27(5):958–969, 2011.

[DSLSV07] A. De Santis, V. Lippiello, B. Siciliano, and L. Villani. Human-robot interaction control using force and vision. In *Advances in Control Theory and Applications*, pages 51–70. Springer, 2007.

[Dur63] N. Durlach. Equalization and cancellation theory of binaural masking-level differences. *The Journal of the Acoustical Society of America*, 35(8):1206–1218, 1963.

[DZD01] R. Duraiswami, D. Zotkin, and L S. Davis. Active speech source localization by a dual coarse-to-fine search. In *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume 5, pages 3309–3312. IEEE, 2001.

[ECR92] B. Espiau, F. Chaumette, and P. Rives. A new approach to visual servoing in robotics. *IEEE Transactions on Robotics and Automation*, 8(3):313–326, 1992.

[EFW+15] J. Even, F. Ferreri, A. Watanabe, Y. Morales, C. Ishi, and N. Hagita. Audio augmented point clouds for applications in robotics. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4846–4851. IEEE, 2015.

[Ell09] DPW Ellis. Gammatone-like spectrograms. *web resource: http://www. ee. columbia. edu/dpwe/resources/matlab/gammatonegram*, 2009.

[EM85] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, 1985.

[EMK+14a] J. Even, Y. Morales, N. Kallakuri, J. Furrer, C. Ishi, and N. Hagita. Mapping sound emitting structures in 3d. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 677–682. IEEE, 2014.

[EMK+14b] J. Even, Y. Morales, N. Kallakuri, C. Ishi, and N. Hagita. Audio ray tracing for position estimation of entities in blind regions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1920–1925. IEEE, 2014.

[EMN14] C. Evers, A H Moore, and P A Naylor. Multiple source localisation in the spherical harmonic domain. In *14th International Workshop on Acoustic Signal Enhancement*, pages 258–262. IEEE, 2014.

[EMN+15] C. Evers, A H Moore, P A Naylor, J. Sheaffer, and B. Rafaely. Bearing-only acoustic tracking of moving speakers for robot audition. In *IEEE International Conference on Digital Signal Processing*, pages 1206–1210. IEEE, 2015.

[EMN16] C. Evers, A H. Moore, and P A. Naylor. Acoustic simultaneous localization and mapping (a-slam) of a moving microphone array and its surrounding speakers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6–10. IEEE, 2016.

[EMS90] B. Espiau, J-P. Merlet, and C. Samson. Force-feedback control and non-contact sensing: a unified approach. In *Proceedings of the 8th CISM-IFTOMM Symposium on Theory and Practice of Robots Manipulators*, 1990.

[ENSHW14] S. Escaida Navarro, M. Schonert, B. Hein, and H. Wörn. 6d proximity servoing for preshaping and haptic exploration using capacitive

tactile proximity sensors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 7–14. IEEE, 2014.

[ESS+09] J. Even, H. Sawada, H. Saruwatari, K. Shikano, and T. Takatani. Semi-blind suppression of internal noise for hands-free robot spoken dialog system. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 658–663. IEEE, 2009.

[FF68] H G. Fisher and S J. Freedman. The role of the pinna in auditory localization. *Journal of Auditory research*, 1968.

[FM89] J. Feddema and O. Mitchell. Vision-guided servoing with feature-based trajectory generation for robots. *IEEE Transactions on Robotics and Automation*, 5(5):691–700, 1989.

[FM02] J. Fredslund and M J. Mataric. A general algorithm for robot formations using local sensing and minimal communication. *IEEE transactions on robotics and automation*, 18(5):837–846, 2002.

[FMG+12] A. Franchi, C. Masone, V. Grabe, M. Ryll, H H Bülthoff, and P. Robuffo Giordano. Modeling and control of uav bearing-formations with bilateral high-level steering. *The International Journal of Robotics Research*, page 0278364912462493, 2012.

[FON+13] K. Furukawa, K. Okutani, K. Nagira, T. Otsuka, K. Itoyama, and H G. Nakadai, K.and Okuno. Noise correlation matrix estimation for improving sound source localization by multirotor uav. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3943–3948. IEEE, 2013.

[GAPFH16] I.D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud. Em algorithms for weighted-data clustering with application to audio-visual scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[GBEH15] I. Gebru, S. Ba, G. Evangelidis, and R. Horaud. Audio-visual speech-turn detection and tracking. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 143–151. Springer, 2015.

[GdM02] J A. Gangloff and M F. de Mathelin. Visual servoing of a 6-dof manipulator for unknown 3-d profile following. *IEEE Transactions on Robotics and Automation*, 18(4):511–520, 2002.

[GESS08] R. Gomez, J. Even, H. Saruwatari, and K. Shikano. Distant talking robust speech recognition using late reflection components of room impulse response. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4581–4584. IEEE, 2008.

[GLF+13] F. Grondin, D. Létourneau, F. Ferland, V. Rousseau, and F. Michaud. The manyears open framework. *Autonomous Robots*, 34(3):217–232, 2013.

[GM90] B R. Glasberg and B. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing research*, 47(1-2):103–138, 1990.

[GM15] F. Grondin and F. Michaud. Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6149–6154. IEEE, 2015.

[GM16] F. Grondin and F. Michaud. Noise mask for tdoa sound source localization of speech on mobile robots in noisy environments. pages 4530–4535, 2016.

[GNN13] R. Gomez, K. Nakamura, and K. Nakadai. Dereverberation robust to speaker's azimuthal orientation in multi-channel human-robot communication. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3439–3444. IEEE, 2013.

[Gre28] G. Green. *An essay on the application of mathematical analysis to the theories of electricity and magnetism.* 1828.

[Gre90] D D. Greenwood. A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America*, 87(6):2592–2605, 1990.

[HAGK03] A A. Handzel, S B. Andersson, M. Gebremichael, and PS Krishnaprasad. A biomimetic apparatus for sound-source localization. In *IEEE Conference on Decision and Control*, volume 6, pages 5879–5884. IEEE, 2003.

[Har99] W M. Hartmann. How we localize sound. *Physics today*, 52:24–29, 1999.

[HC05] S Haykin and Z. Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[HCM83] R Hausler, S Colburn, and E Marr. Sound localization in subjects with impaired hearing, spatial-discrimination and interaural-discrimination tests. *Acta Oto-Laryngologica*, pages 1–62, 1983.

[HCM94] G. Hager, W-C Chang, and S. Morse. Robot feedback control based on stereo vision: Towards calibration-free hand-eye coordination. In *IEEE International Conference on Robotics and Automation*, pages 2850–2856. IEEE, 1994.

[HCW+11] J-S.g Hu, C-Y. Chan, C-K. Wang, M-T. Lee, and C-Y. Kuo. Simultaneous localization of a mobile robot and multiple sound sources using a microphone array. *Advanced Robotics*, 25(1-2):135–152, 2011.

[HD69] RM Hershkowitz and NI Durlach. Interaural time and amplitude jnds for a 500-hz tone. *The Journal of the Acoustical Society of America*, 46(6B):1464–1467, 1969.

[HDE98] R. Horaud, F. Dornaika, and B. Espiau. Visually guided object grasping. *IEEE Transactions on Robotics and Automation*, 14(4):525–532, 1998.

[Her86] D. Hertz. Time delay estimation by combining efficient algorithms and generalized cross-correlation methods. *IEEE transactions on acoustics, speech, and signal processing*, 34(1):1–7, 1986.

[HIA98] K. Hosoda, K. Igarashi, and M. Asada. Adaptive hybrid control for visual and force servoing in an unknown environment. *IEEE Robotics & Automation Magazine*, 5(4):39–43, 1998.

[HLSVL06] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda. Sound localization for humanoid robots-building audio-motor maps based on the hrtf. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1170–1176. IEEE, 2006.

[HNK+97] S. Hashimoto, S Narita, H Kasahara, A. Takanishi, S Sugano, K Shirai, T Kobayashi, H Takanobu, T Kurata, K Fujiwara, et al. Humanoid robot-development of an information assistant robot hadaly. In *IEEE International Workshop on Robot and Human Communication*, pages 106–111. IEEE, 1997.

[HOS95] J. Huang, N. Ohnishi, and N. Sugie. A biomimetic system for localization and separation of multiple sound sources. *IEEE transactions on Instrumentation and Measurement*, 44(3):733–738, 1995.

[HOS97] J. Huang, N. Ohnishi, and N. Sugie. Building ears for robots: sound localization and separation. *Artificial Life and Robotics*, 1(4):157–163, 1997.

[HST+99] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Ohnishi, and N. Sugie. A model-based sound localization system and its application to robot navigation. *Robotics and Autonomous Systems*, 27(4):199–209, 1999.

[HSTO97] J. Huang, T. Supaongprapa, I. Terakura, and N. Ohnishi, N.and Sugie. Mobile robot and sound localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 683–689. IEEE, 1997.

[HT02] L Harry and V. Trees. Optimum array processing: part iv of detection, estimation, and modulation theory. *New York: John Wiler & Sons*, 2002.

[HTOO11] Y. Hirasawa, T. Takahashi, T. Ogata, and H G Okuno. Robot with two ears listens to more than two simultaneous utterances by exploiting harmonic structures. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 348–358. Springer, 2011.

[HVRVO98] P M Hofman, J. GA Van Riswick, and A J Van Opstal. Relearning sound localization with new ears. *Nature neuroscience*, 1(5):417–421, 1998.

[HWA08] RW Hill, GA Wyse, and M Anderson. Animal physiology. masschasusett sinauer ass. *Inc. Pub. Sounderland*, 2008.

[HYT+12] Y. Hirasawa, N. Yasuraoka, T. Takahashi, T. Ogata, and H G Okuno. A gmm sound source model for blind speech separation in underdetermined conditions. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 446–453. Springer, 2012.

[IKSM05] A. Ito, T. Kanayama, M. Suzuki, and S. Makino. Internal noise suppression for speech recognition by small robots. In *INTERSPEECH*, pages 2685–2688, 2005.

[INA+11] G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai. Assessment of general applicability of ego noise estimation. In *IEEE International Conference on Robotics and Automation*, pages 3517–3522. IEEE, 2011.

[INR+09] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. Imura. Ego noise suppression of a robot using template subtraction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 199–204. IEEE, 2009.

[Iri95] R E Irie. *Robust sound localization: An application of an auditory perception system for a humanoid robot.* PhD thesis, Massachusetts Institute Of Technology, 1995.

[JAPGH12] M. Janvier, X. Alameda-Pineda, L. Girinz, and R. Horaud. Sound-event recognition with a companion humanoid. In *IEEE-RAS International Conference on Humanoid Robots*, pages 104–111. IEEE, 2012.

[JD92] D H. Johnson and D E. Dudgeon. *Array signal processing: concepts and techniques.* Simon & Schuster, 1992.

[Jef48] L A. Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35, 1948.

[Joh72] P. Johannesma. The pre-response stimulus ensemble of neurons in the cochlear nucleus. In *Symposium on Hearing Theory*, pages 58–69. IPO Eindhoven, Holland, 1972.

[JSY98] P X. Joris, P. Smith, and T C. Yin. Coincidence detection in the auditory system: 50 years after jeffress. *Neuron*, 21(6):1235–1238, 1998.

[JY95] P X. Joris and T C Yin. Envelope coding in the lateral superior olive. i. sensitivity to interaural time differences. *Journal of Neurophysiology*, 73(3):1043–1062, 1995.

[KB08] L. Kneip and C. Baumann. Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis. *The Journal of the Acoustical Society of America*, 124(5):3108–3119, 2008.

[KBGAP+15] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud. A variational em algorithm for the separation of moving sound sources. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–5. IEEE, 2015.

[KC76] C.H Knapp and G.C Carter. The generalized correlation method for estimation of time delay. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 24(4):320–327, 1976.

[KCCL08] C-T. Kim, T-Y. Choi, B. Choi, and J-J. Lee. Robust estimation of sound direction for robot interface. In *IEEE International Conference on Robotics and Automation*, pages 3475–3480. IEEE, 2008.

[KDV16] M. Krekovic, I. Dokmanic, and M. Vetterli. Echoslam: Simultaneous localization and mapping with acoustic echoes. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, number EPFL-CONF-215300, 2016.

[KFKI10] M. Kumon, K. Fukushima, S. Kunimatsu, and M. Ishitobi. Motion planning based on simultaneous perturbation stochastic approximation for mobile auditory robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 431–436. IEEE, 2010.

[KGD+03] A. Krupa, J. Gangloff, C. Doignon, M F. de Mathelin, G. Morel, J. Leroy, L. Soler, and J. Marescaux. Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. *IEEE transactions on robotics and automation*, 19(5):842–853, 2003.

[KGD13] A. Kalmbach, Y. Girdhar, and G. Dudek. Unsupervised environment recognition and modeling using sound sensing. In *IEEE International Conference on Robotics and Automation*, pages 2699–2704. IEEE, 2013.

[KH16] L. Kumar and R M Hegde. Near-field acoustic source localization and beamforming in spherical harmonics domain. *IEEE Transactions on Signal Processing*, 64(13):3351–3361, 2016.

[KJHJ05] M. Karasalo, L-M Johansson, X. Hu, and K H. Johansson. Multi-robot terrain servoing with proximity sensors. In *IEEE International Conference on Robotics and Automation*, pages 2791–2796, 2005.

[KKK+15] P. Kriengkomol, K. Kamiyama, M. Kojima, M. Horade, Y. Mae, and T. Arai. Hammering sound analysis for infrastructure inspection by leg robot. In *IEEE International Conference on Robotics and Biomimetics*, pages 887–892. IEEE, 2015.

[KMIH13] J. Kallakuri, N.and Even, Y. Morales, C. Ishi, and N. Hagita. Probabilistic approach for building auditory maps with a mobile microphone array. In *IEEE International Conference on Robotics and Automation*, pages 2270–2275. IEEE, 2013.

[KMOO11] U-H Kim, T. Mizumoto, T. Ogata, and H G Okuno. Improvement of speaker localization by considering multipath interference of sound wave for binaural robot audition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2910–2915. IEEE, 2011.

[KN11] M. Kumon and Y. Noda. Active soft pinnae for robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 112–117. IEEE, 2011.

[KND06] F. Keyrouz, Y. Naous, and K. Diepold. A new method for binaural 3-d localization based on hrtfs. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages V–V. IEEE, 2006.

[KNM15] I. Kossyk, M. Neumann, and Z-C. Marton. Binaural bearing only tracking of stationary sound sources in reverberant environment. In *IEEE-RAS 15th International Conference on Humanoid Robots*, pages 53–60. IEEE, 2015.

[KOK+74] I. Kato, S. Ohteru, H. Kobayashi, K. Shirai, and A. Uchiyama. Information-power machine with senses and limbs. In *On Theory and Practice of Robots and Manipulators*, pages 11–24. Springer, 1974.

[KOS+87] I. Kato, S. Ohteru, K. Shirai, T. Matsushima, S. Narita, S. Sugano, T. Kobayashi, and E. Fujisawa. The robot musician 'wabot-2'(waseda robot-2). *Robotics*, 3(2):143–155, 1987.

[KPTK12] H M Kondo, D. Pressnitzer, I. Toshima, and M. Kashino. Effects of self-motion on auditory scene analysis. *Proceedings of the National Academy of Sciences*, 109(17):6775–6780, 2012.

[KSK+05] M. Kumon, T. Shimoda, R. Kohzawa, I. Mizumoto, and Z. Iwai. Audio servo for robotic systems with pinnae. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1881–1886. IEEE, 2005.

[KSM+03] M. Kumon, T. Sugawara, K. Miike, I. Mizumoto, and Z. Iwai. Adaptive audio servo for multirate robot systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 1, pages 182–187. IEEE, 2003.

[KSN15] R. Kojima, O. Sugiyama, and K. Nakadai. Scene understanding based on sound and text information for a cooking support robot. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 665–674. Springer, 2015.

[KUKH03] M. Kato, H. Uematsu, M. Kashino, and T. Hirahara. The effect of head motion on the accuracy of sound localization. *Acoustical Science and Technology*, 24(5):315–317, 2003.

[LC08] Y. Lu and M. Cooke. Strategic listener movement in a model of auditory distance perception. *The Journal of the Acoustical Society of America*, 123(5):3726–3726, 2008.

[LC10] Ya-C Lu and M. Cooke. Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1793–1805, 2010.

[LC11] Y. Lu and M. Cooke. Motion strategies for binaural localisation of speech sources in azimuth and distance by artificial listeners. *Speech Communication*, 53(5):622–642, 2011.

[LDW91] J J Leonard and H F Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *IEEE Transactions on robotics and Automation*, 7(3):376–382, 1991.

[LHC90] J. Loomis, C. Hebert, and J. Cicinelli. Active localization of virtual sounds. *The Journal of the Acoustical Society of America*, 88(4):1757–1764, 1990.

[LLY11] X. Li, H. Liu, and X. Yang. Sound source localization for mobile robot based on time difference feature and space grid matching. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2879–2886. IEEE, 2011.

[LPCS10] M. Liu, C. Pradalier, Q. Chen, and R. Siegwart. A bearing-only 2d/3d-homing method under a visual servoing framework. In *IEEE International Conference on Robotics and Automation*, pages 4062–4067, 2010.

[LSHR13] Q. Li, C. Schürmann, R. Haschke, and H J Ritter. A control framework for tactile servoing. In *Robotics: Science and systems*, 2013.

[LSWL12] X. Li, M. Shen, W. Wang, and H. Liu. Real-time sound source localization for a mobile robot based on the guided spectral-temporal position method. *International Journal of Advanced Robotic Systems*, 9, 2012.

[MAB11] E. Martinson, T. Apker, and M. Bugajska. Optimizing a reconfigurable robotic microphone array. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 125–130. IEEE, 2011.

[MBM15] N. Ma, G J Brown, and T. May. Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions. *INTERSPEECH*, pages 3302–3306, 2015.

[MC02] E. Marchand and F. Chaumette. Virtual visual servoing: a framework for real-time augmented reality. In *Computer Graphics Forum*, volume 21, pages 289–297. Wiley Online Library, 2002.

[MEN+15] A H Moore, C. Evers, P A Naylor, D L Alon, and B Rafaely. Direction of arrival estimation using pseudo-intensity vectors with direct-path dominance test. In *23rd European Signal Processing Conference*, pages 2296–2300. IEEE, 2015.

[MEW09] J C. Murray, H. Erwin, and S. Wermter. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. *Neural Networks*, 22(2):173–189, 2009.

[MG90] J C. Middlebrooks and D M. Green. Directional dependence of interaural envelope delays. *The Journal of the Acoustical Society of America*, 87(5):2149–2162, 1990.

[MG91] J C. Middlebrooks and D M. Green. Sound localization by human listeners. *Annual review of psychology*, 42(1):135–159, 1991.

[MG03] D. McAlpine and B. Grothe. Sound localization and delay lines–do mammals fit the model? *Trends in neurosciences*, 26(7):347–350, 2003.

[MH07] N. Metni and T. Hamel. A uav for bridge inspection: Visual servoing control law with orientation limits. *Automation in construction*, 17(1):3–10, 2007.

[MI68] P M Morse and K U Ingard. *Theoretical acoustics*. Princeton university press, 1968.

[Mil72] A W. Mills. Auditory localization(binaural acoustic field sampling, head movement and echo effect in auditory localization of sound sources position, distance and orientation). *Foundations of modern auditory theory.*, 2:303–348, 1972.

[MKC08] R. Mebarki, A. Krupa, and F. Chaumette. Image moments-based ultrasound visual servoing. In *IEEE International Conference on Robotics and Automation*, pages 113–119. IEEE, 2008.

[MM90] J. Makous and J. Middlebrooks. Two-dimensional sound localization by human listeners. *The journal of the Acoustical Society of America*, 87(5):2188–2200, 1990.

[MMB15] T. May, N. Ma, and G J Brown. Robust localisation of multiple speakers exploiting head movements and multi-conditional training of binaural cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2679–2683. IEEE, 2015.

[MMBB15] N. Ma, R. Marxer, J. Barker, and G J Brown. Exploiting synchrony spectra and deep neural networks for noise-robust automatic speech recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 490–495. IEEE, 2015.

[MMR10] E. Malis, Y. Mezouar, and P. Rives. Robustness of image-based visual servoing with a calibrated camera in the presence of uncertainties in the three-dimensional structure. *IEEE Transactions on Robotics*, 26(1):112–120, 2010.

[MMWB15] N. Ma, T. May, H. Wierstorf, and G. Brown. A machine-hearing system exploiting head movements for binaural sound localisation in reverberant conditions. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2699–2703. IEEE, 2015.

[MOD89] B. Moore, S R. Oldfield, and G J. Dooley. Detection and discrimination of spectral peaks and notches at 1 and 8 khz. *The Journal of the Acoustical Society of America*, 85(2):820–836, 1989.

[MP10] I. Marković and I. Petrović. Speaker localization and tracking with a microphone array on a mobile robot using von mises distribution and particle filtering. *Robotics and Autonomous Systems*, 58(11):1185–1196, 2010.

[MPS05] A. Martinelli, F. Pont, and R. Siegwart. Multi-robot localization using relative observations. In *IEEE international conference on robotics and automation*, pages 2797–2802. IEEE, 2005.

[MR+93] H. Michel, P. Rives, et al. Singularities in the determination of the situation of a robot effector from the perspective view of 3 points. *INRIA Research Report, Tech. Rep. 1850*, 1993.

[MvdPK11] T. May, S. van de Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on audio, speech, and language processing*, 19(1):1–13, 2011.

[NALRL14] D. Navarro-Alarcon, Y-H. Liu, J. Romero, and P. Li. On the visual deformation servoing of compliant objects: Uncalibrated control methods and experiments. *The International Journal of Robotics Research*, 33(11):1462–1480, 2014.

[NCVC16] V Q. Nguyen, F. Colas, E. Vincent, and F. Charpillet. Localizing an intermittent and moving sound source using a mobile robot. In *IEEE International Conference on Intelligent Robots and Systems*, 2016.

[NGN13] K. Nakamura, R. Gomez, and K. Nakadai. Real-time super-resolution three-dimensional sound source localization for robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3949–3954. IEEE, 2013.

[NHOK02] K. Nakadai, K. Hidai, H G Okuno, and H. Kitano. Real-time speaker localization and speech separation by audio-visual integration. In *IEEE International Conference on Robotics and Automation,*, volume 1, pages 1043–1049. IEEE, 2002.

[Nie92] S. Nielsen. Auditory distance perception in different rooms. In *Audio Engineering Society Convention 92*. Audio Engineering Society, 1992.

[NIYO15] I. Nishimuta, K. Itoyama, K. Yoshii, and Hiroshi G Okuno. Toward a quizmaster robot for speech-based multiparty interaction. *Advanced Robotics*, 29(18):1205–1219, 2015.

[NKD+09] H. Nakajima, K. Kikuchi, T. Daigo, Y. Kaneda, K. Nakadai, and Y. Hasegawa. Real-time sound source orientation estimation using a 96 channel microphone array. In *IEEE Int. Conf. on Intelligent Robots and Systems*, pages 676–683, 2009.

[NMN15] K. Nakadai, T. Mizumoto, and K. Nakamura. Robot-audition-based human-machine interface for a car. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6129–6136. IEEE, 2015.

[NMOK03] K. Nakadai, D. Matsuura, H G Okuno, and H. Kitano. Applying scattering theory to robot audition system: Robust sound source localization and extraction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 1147–1152. IEEE, 2003.

[NMOT04] K. Nakadai, D. Matsuura, H G Okuno, and H. Tsujino. Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. *Speech Communication*, 44(1):97–112, 2004.

[NMS02] L. Natale, G. Metta, and G. Sandini. Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head. *Robotics and Autonomous Systems*, 39(2):87–106, 2002.

[NNA+09] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino. Intelligent sound source localization for dynamic environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 664–669. IEEE, 2009.

[NNHT09] K. Nakadai, H. Nakajima, Y. Hasegawa, and H. Tsujino. Sound source separation of moving speakers for robot audition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3685–3688. IEEE, 2009.

[NNI12] K. Nakamura, K. Nakadai, and G. Ince. Real-time super-resolution sound source localization for robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 694–699. IEEE, 2012.

[NNIH10] K. Nakadai, H. Nakajima, G. Ince, and Y. Hasegawa. Sound source separation and automatic speech recognition for moving sources. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 976–981. IEEE, 2010.

[NNN+06] Y. Nishimura, M. Nakano, K. Nakadai, H. Tsujino, and M. Ishizuka. Speech recognition for a robot under its motor noises by selective application of missing feature theory and mllr. In *INTERSPEECH*, pages 53–58, 2006.

[NOK00a] K. Nakadai, H G Okuno, and H. Kitano. Humanoid active audition system improved by the cover acoustics. In *Pacific Rim International Conference on Artificial Intelligence*, pages 544–554. Springer, 2000.

[NOK00b] T. Nakadai, K.and Lourens, H G. Okuno, and H. Kitano. Active audition for humanoid. In *AAAI/IAAI*, pages 832–839, 2000.

[NOK01] K. Nakadai, H G Okuno, and H. Kitano. Epipolar geometry based sound localization and extraction for humanoid audition. In

IEEE/RSJ International Conference on Intelligent Robots and Systems, volume 3, pages 1395–1401. IEEE, 2001.

[NOK02] K. Nakadai, H G. Okuno, and H. Kitano. Auditory fovea based speech separation and its application to dialog system. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 1320–1325. IEEE, 2002.

[NSN15] K. Nakamura, L. Sinapayen, and K. Nakadai. Interactive sound source localization using robot audition for tablet devices. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6137–6142. IEEE, 2015.

[NTO+10] K. Nakadai, Toru Takahashi, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. Design and implementation of robot audition system'hark'—open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[NTOO12] K. Nagira, T. Takahashi, T. Ogata, and H G Okuno. Complex extension of infinite sparse factor analysis for blind speech separation. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 388–396. Springer, 2012.

[OBI+15] M. Ohkita, Y. Bando, Y. Ikemiya, K.i Itoyama, and K. Yoshii. Audio-visual beat tracking based on a state-space model for a music robot dancing with humans. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5555–5560. IEEE, 2015.

[OI06] M. Otani and S. Ise. Fast calculation system specialized for head-related transfer function based on boundary element method. *The Journal of the Acoustical Society of America*, 119(5):2589–2598, 2006.

[ON15] H G Okuno and K. Nakadai. Robot audition: Its rise and perspectives. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5610–5614. IEEE, 2015.

[ONM+14] T. Ohata, K. Nakamura, T. Mizumoto, T. Taiki, and K. Nakadai. Improvement in outdoor sound source detection using a quadrotor-embedded microphone array. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1902–1907. IEEE, 2014.

[ONOO11] T. Otsuka, K. Nakadai, T. Ogata, and H G Okuno. Bayesian extension of music for sound source localization and tracking. In *INTERSPEECH*, pages 3109–3112, 2011.

[ONT+11] Takuma Otsuka, Kazuhiro Nakadai, Toru Takahashi, Tetsuya Ogata, and Hiroshi G Okuno. Real-time audio-to-score alignment using particle filter for coplayer music robots. *EURASIP Journal on Advances in Signal Processing*, 2011:2, 2011.

[OOKN04] H G Okuno, T. Ogata, K. Komatani, and K. Nakadai. Computational auditory scene analysis and its application to robot audition. In *International Conference on Informatics Research for Development of Knowledge Society Infrastructure*, pages 73–80. IEEE, 2004.

[OP86] S R. Oldfield and S. Parker. Acuity of sound localisation: a topography of auditory space. iii. monaural hearing conditions. *Perception*, 15(1):67–81, 1986.

[OSC00] A M. Okamura, N Smaby, and M R. Cutkosky. An overview of dexterous manipulation. In *IEEE International Conference on Robotics and Automation*, volume 1, pages 255–262. IEEE, 2000.

[OYNN12] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai. Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3288–3293. IEEE, 2012.

[P+81] A D Pierce et al. *Acoustics: an introduction to its physical principles and applications*, volume 20. McGraw-Hill New York, 1981.

[PAG95] R D. Patterson, M H. Allerhand, and C. Giguere. Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, 98(4):1890–1894, 1995.

[PB93] PW. Poon and J F. Brugge. Sensitivity of auditory nerve fibers to spectral notches. *Journal of neurophysiology*, 70(2):655–666, 1993.

[PBD+14] A. Portello, G. Bustamante, P. Danès, J. Piat, and J. Manhès. Active localization of an intermittent sound source from a moving binaural sensor. European Acoustics Association, 2014.

[PBDM14] A. Portello, G. Bustamante, P. Danes, and A. Mifsud. Localization of multiple sources from a binaural head in a known noisy environment. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3168–3174. IEEE, 2014.

[PDA12] A. Portello, P. Danes, and S. Argentieri. Active binaural localization of intermittent moving sources in the presence of false measurements. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3294–3299. IEEE, 2012.

[PM86] R. D. Patterson and Brian C. J. Moore. Auditory filters and excitation patterns as representations of frequency resolution. *Frequency selectivity in hearing*, pages 123–177, 1986.

[PN97a] S. Perrett and W. Noble. The contribution of head motion cues to localization of low-pass noise. *Perception & psychophysics*, 59(7):1018–1026, 1997.

[PN97b] S. Perrett and W. Noble. The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America*, 102(4):2325–2332, 1997.

[Pop08] L C Populin. Human sound localization: measurements in untrained, head-unrestrained subjects using gaze as a pointer. *Experimental brain research*, 190(1):11–30, 2008.

[PPL90] D. Pearson, S U. Pillai, and Y. Lee. An algorithm for near-optimal placement of sensor elements. *IEEE Transactions on Information Theory*, 36(6):1280–1284, 1990.

[PRH+92] R D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. *Auditory physiology and perception*, 83:429–446, 1992.

[Ray96] J W S B Rayleigh. *The theory of sound*, volume 2. Macmillan, 1896.

[Ray07] L. Rayleigh. Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

[RCE90] P. Rives, F. Chaumette, and B. Espiau. Visual servoing based on a task function approach. In *Experimental Robotics I*, pages 412–428. Springer, 1990.

[RDY05] V. Raykar, R. Duraiswami, and B. Yegnanarayana. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *The Journal of the Acoustical Society of America*, 118(1):364–374, 2005.

[RF04] Y. Rui and D. Florencio. Time delay estimation in the presence of correlated noise and reverberation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing,*, volume 2, pages ii–133. IEEE, 2004.

[RIJG08] T. Rodemann, G. Ince, F. Joublin, and C. Goerick. Using binaural and spectral cues for azimuth and elevation localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2185–2190. IEEE, 2008.

[RO98] D Rosenthal and H G Okuno. *Computational auditory scene analysis.* Lawrence Erlbaum Associates Publishers, 1998.

[Rod10] T. Rodemann. A study on distance estimation in binaural sound localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 425–430. IEEE, 2010.

[Rot71] P R. Roth. Effective measurements using digital signal analysis. *IEEE spectrum*, 4(8):62–70, 1971.

[RVE10] M. Raspaud, H. Viste, and G. Evangelista. Binaural source localization by joint estimation of ild and itd. *IEEE Transaction on Audio, Speech, and Language Processing*, 18(1):68–77, 2010.

[RWE98] M Rucci, J Wray, and GM Edelman. Spatial localization and the refinement of orienting behavior: What can be learned from the barn owl? In *IEEE International Symposium on Intelligent Control*, pages 253–258. IEEE, 1998.

[SAA91] C. Samson and K. Ait-Abderrahim. Feedback control of a nonholonomic wheeled cart in cartesian space. In *IEEE International Conference on Robotics and Automation*, pages 1136–1141. IEEE, 1991.

[Saa96] H. Saarnisaari. Ml time delay estimation in a multipath channel. In *IEEE 4th International Symposium on Spread Spectrum Techniques and Applications*, volume 3, pages 1007–1011. IEEE, 1996.

[SC09] J. Schnupp and C. Carr. On hearing with more than one ear: lessons from evolution. *Nature neuroscience*, 12(6):692–697, 2009.

[Sch86] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[SCSK00] Barbara G Shinn-Cunningham, Scott Santarelli, and Norbert Kopco. Tori of confusion: Binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 107(3):1627–1636, 2000.

[SELB91] C. Samson, B. Espiau, and M. Le Borgne. *Robot control: the task function approach*. Oxford University Press, 1991.

[SG62] S. Stevens and M. Guirao. Loudness, reciprocality, and partition scales. *The Journal of the Acoustical Society of America*, 34(9B):1466–1471, 1962.

[She82] C. Sheeline. *An investigation of the effects of direct and reverberant signal interaction on auditory distance perception*. PhD thesis, Stanford University, 1982.

[SII96] T. Shibata, K. Inoue, and R. Irie. Emotional robot for intelligent system-artificial emotional creature project. In *IEEE International Workshop on Robot and Human Communication*, pages 466–471. IEEE, 1996.

[SJC78] R M. Stern Jr and H S. Colburn. Theory of binaural interaction based on auditory-nerve data. iv. a model for subjective lateral position. *The Journal of the Acoustical Society of America*, 64(1):127–140, 1978.

[SJHS15] TM Sreejith, PK Joshin, S Harshavardhan, and TV Sreenivas. Tde sign based homing algorithm for sound source tracking using a y-shaped microphone array. In *23rd European Signal Processing Conference (EUSIPCO)*, pages 1202–1206. IEEE, 2015.

[SKHO08] S. Shiramatsu, K. Komatani, K. Hasida, and H G Ogata, T.and Okuno. A game-theoretic model of referential coherence and its empirical verification using large japanese and english corpora. *ACM Transactions on Speech and Language Processing*, 5(3):6, 2008.

[SKM06] Y. Sasaki, S. Kagami, and H. Mizoguchi. Multiple sound source mapping for a mobile robot by self-motion triangulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 380–385. IEEE, 2006.

[SKT+12] Y. Sasaki, M. Kabasawa, S. Thompson, S. Kagami, and K. Oro. Spherical microphone array for spatial sound localization for a mobile robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 713–718. IEEE, 2012.

[Sla88] M. Slaney. *Lyon's cochlear model*, volume 13. Citeseer, 1988.

[SM15] R V. Sharan and T J. Moir. Cochleagram image feature for improved robustness in sound recognition. In *IEEE International Conference on Digital Signal Processing*, pages 441–444. IEEE, 2015.

[SN36] S. Smith Stevens and E B. Newman. The localization of actual sources of sound. *The American Journal of Psychology*, 48(2):297–306, 1936.

[Str77] J W Strutt. *The theory of sound*, volume 1. 1877.

[SUW15] D. Su, N. Ulapane, and B.i Wijerathna. An acoustic sensor based novel method for 2d localization of a robot in a structured environment. In *IEEE 10th Conference on Industrial Electronics and Applications*, pages 2024–2029. IEEE, 2015.

[SVN37] S S. Stevens, J. Volkmann, and E B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

[SWK+14] C. Schymura, T. Walther, D. Kolossa, N. Ma, and G J Brown. Binaural sound source localisation using a bayesian-network-based blackboard system and hypothesis-driven feedback. *Forum Acusticum*, 2014.

[SZS94] P. Sikka, H. Zhang, and S. Sutphen. Tactile servo: Control of touch-driven robot motion. In *Experimental Robotics III*, pages 219–233. Springer, 1994.

[TA06] I. Toshima and S. Aoki. The effect of head movement on sound localization in an acoustical telepresence robot: Telehead. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 872–877. IEEE, 2006.

[TAM+02] S. Takane, D. Arai, T. Miyajima, K. Watanabe, Y. Suzuki, and T. Sone. A database of head-related transfer functions in whole directions on upper hemisphere. *Acoustical science and technology*, 23(3):160–162, 2002.

[TM12] C. Teuliere and E. Marchand. Direct 3d servoing using dense depth maps. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1741–1746. IEEE, 2012.

[TMR67] W. Thurlow, J. Mangels, and P. Runge. Head movements during sound localization. *The Journal of the Acoustical society of America*, 42(2):489–493, 1967.

[TNT+12] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and H G Okuno. Efficient blind dereverberation and echo cancellation based on independent component analysis for actual acoustic signals. *Neural computation*, 24(1):234–272, 2012.

[TOO14] T. Tasaki, T. Ogata, and H G Okuno. The interaction between a robot and multiple people based on spatially mapping of friendliness and motion parameters. *Advanced Robotics*, 28(1):39–51, 2014.

[TR13] V. Tourbabin and B. Rafaely. Theoretical framework for the design of microphone arrays for robot audition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4290–4294, 2013.

[TR14] V. Tourbabin and B. Rafaely. Theoretical framework for the optimization of microphone array configuration for humanoid robot audition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1803–1814, 2014.

[TR15] V. Tourbabin and B. Rafaely. Direction of arrival estimation using microphone array processing for moving humanoid robots. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(11):2046–2058, 2015.

[Tre02] H. Trees. Optimum array processing, detection, estimation and modulation part iv. *IEEE Trans, John Wiley and Sons, Inc., New York*, 2002.

[TSM+04] T M. Talavage, M I. Sereno, J R. Melcher, P J. Ledden, B R. Rosen, and A M. Dale. Tonotopic organization in human auditory cortex

revealed by progressions of frequency sensitivity. *Journal of neurophysiology*, 91(3):1282–1296, 2004.

[TTLJ15] N. Thomsen, Z-H Tan, B. Lindberg, and S. Jensen. A heuristic approach for a social robot to navigate to a person based on audio and range information. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5884–5890. IEEE, 2015.

[TY02] D J. Tollin and T. Yin. The coding of spatial location by single units in the lateral superior olive of the cat. i. spatial receptive fields in azimuth. *The Journal of neuroscience*, 22(4):1454–1467, 2002.

[TY03] D J. Tollin and T. Yin. Spectral cues explain illusory elevation effects with stereo sounds in cats. *Journal of neurophysiology*, 90(1):525–530, 2003.

[VBW60] G. Von Békésy and E G. Wever. *Experiments in hearing*, volume 8. McGraw-Hill New York, 1960.

[Ves09] S. Vesa. Binaural sound source distance learning in rooms. *IEEE Transactions on audio, speech, and language processing*, 17(8):1498–1507, 2009.

[vHW20] EM von Hornbostel and M. Wertheimer. Über die wahrnehmung der schallrichtung [on the perception of the direction of sound]. *Sitzungsberichte der preussischen Akademie der Wissenschaften*, 388:396, 1920.

[VMHR04] J-M. Valin, F. Michaud, B. Hadjou, and J. Rouat. Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach. In *IEEE International Conference on Robotics and Automation*, volume 1, pages 1033–1038. IEEE, 2004.

[VMR06] J-M. Valin, F.s Michaud, and J. Rouat. Robust 3d localization and tracking of sound sources using beamforming and particle filtering. In *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 4, pages IV–IV. IEEE, 2006.

[VMR07] J-M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems*, 55(3):216–228, 2007.

[VMRL03] J-M. Valin, F. Michaud, J. Rouat, and D. Létourneau. Robust sound source localization using a microphone array on a mobile robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 1228–1233. IEEE, 2003.

[VSC15] E. Vincent, A. Sini, and F. Charpillet. Audio source localization by optimal control of a mobile robot. In *IEEE International Conference on Acoustics, Speech and Signal Processing 5*, pages 5630–5634. IEEE, 2015.

[VVB88] B D. Van Veen and K M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP magazine*, 5(2):4–24, 1988.

[VVGVO04] J. Vliegen, T J Van Grootel, and A J Van Opstal. Dynamic sound localization during rapid eye-head gaze shifts. *The Journal of neuroscience*, 24(42):9291–9302, 2004.

[VVO04] J. Vliegen and A J Van Opstal. The influence of duration and level on human sound localization. *The Journal of the Acoustical Society of America*, 115(4):1705–1713, 2004.

[WAKW93] E M. Wenzel, M. Arruda, D J. Kistler, and F L. Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 1993.

[Wal39] H. Wallach. On sound localization. *The Journal of the Acoustical Society of America*, 10(4):270–274, 1939.

[Wal40] H. Wallach. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339, 1940.

[War68] RM Warren. Vocal compensation for change in distance. In *Proceedings of the 6th International Congress of Acoustics (Tokyo), A*, pages 61–64, 1968.

[Web93] B. Webb. Modeling biological behaviour or" dumb animals and stupid robots". *DAI RESEARCH PAPER*, 1993.

[Web95] B. Webb. Using robots to model animals: a cricket test. *Robotics and autonomous systems*, 16(2):117–134, 1995.

[Wev49] E G. Wever. Theory of hearing. 1949.

[WGS11] H. Wierstorf, M. Geier, and S. Spors. A free database of head related impulse response measurements in the horizontal plane with multiple distances. In *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

[WHW74] D. Wright, J H. Hebrank, and B. Wilson. Pinna reflections as cues for localization. *The Journal of the Acoustical Society of America*, 56(3):957–962, 1974.

[WIA04] Q H. Wang, T. Ivanov, and P. Aarabi. Acoustic robot navigation using distributed microphone arrays. *Information Fusion*, 5(2):131–140, 2004.

[Wie47] F. Wiener. On the diffraction of a progressive sound wave by the human head. *The Journal of the Acoustical Society of America*, 19(1):143–146, 1947.

[WK89] F L. Wightman and D J. Kistler. Headphone simulation of free-field listening. ii: Psychophysical validation. *The Journal of the Acoustical Society of America*, 85(2):868–878, 1989.

[WK92] F L. Wightman and D J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *The Journal of the Acoustical Society of America*, 91(3):1648–1661, 1992.

[WK99] F. L Wightman and D J Kistler. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America*, 105(5):2841–2853, 1999.

[WL03] D. Ward and R. Lehmann, E.and Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Transactions on speech and audio processing*, 11(6):826–836, 2003.

[WNR49] H. Wallach, E B. Newman, and M R. Rosenzweig. A precedence effect in sound localization. *The Journal of the Acoustical Society of America*, 21(4):468–468, 1949.

[WS08] R. Wistort and J R. Smith. Electric field servoing for robotic manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 494–499. IEEE, 2008.

[WSN87] L. Weiss, A. Sanderson, and C. Neuman. Dynamic sensor-based control of robots with visual feedback. *IEEE Journal on Robotics and Automation*, 3(5):404–417, 1987.

[WW12] J. Woodruff and D. Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Transaction on Audio, Speech, and Language Processing*, 20(5):1503–1512, 2012.

[YA95] N. Yamasaki and Y. Anzai. Active interface for human-robot interaction. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 3103–3109. IEEE, 1995.

[YAZ12a] K. Youssef, S. Argentieri, and J-L. Zarader. A binaural sound source localization method using auditive cues and vision. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 217–220. IEEE, 2012.

[YAZ12b] K. Youssef, S. Argentieri, and J-L. Zarader. Towards a systematic study of binaural cues. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1004–1009. IEEE, 2012.

[YAZ13] K. Youssef, S. Argentieri, and J-L Zarader. A learning-based approach to robust binaural sound localization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2927–2932. IEEE, 2013.

[YBA+11] K. Youssef, B. Breteau, S. Argentieri, J-L Zarader, and Z. Wang. Approaches for automatic speaker recognition in a binaural humanoid context. In *ESANN*, 2011.

[YN01] S. Yu and B J. Nelson. Autonomous injection of biological cells using visual servoing. In *Experimental Robotics VII*, pages 169–178. Springer, 2001.

[YNT+07] K. Yoshii, K. Nakadai, T. Torii, Y. Hasegawa, H. Tsujino, K. Komatani, T. Ogata, and H G. Okuno. A biped robot that keeps steps in time with musical beats while listening to music with its own ears. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1743–1750. IEEE, 2007.

[Yos94] W A. Yost. *Fundamentals of hearing: An introduction* . Academic Press, 1994.

[YSD+12] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann. Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition. *IEEE Signal Processing Magazine*, 29(6):114–126, 2012.

[YTK+11] N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, and H G Okuno. Environmental sound recognition for robot audition using matching-pursuit. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 1–10. Springer, 2011.

[ZB10] X. Zhu and J R Buck. Designing nonuniform linear arrays to maximize mutual information for bearing estimationa). *The Journal of the Acoustical Society of America*, 128(5):2926–2939, 2010.

[ZBB05] P. Zahorik, D S Brungart, and A W Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91(3):409–420, 2005.

[ZDD04] D N. Zotkin, R. Duraiswami, and L S. Davis. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia*, 6(4):553–564, 2004.

[ZVVO04] M P. Zwiers, H. Versnel, and A J. Van Opstal. Involvement of monkey inferior colliculus in spatial hearing. *The Journal of neuroscience*, 24(17):4145–4156, 2004.

[ZX14] X-L. Zhong and B-S. Xie. Head-related transfer functions and virtual auditory display. *Soundscape Semiotics-Localization and Categorization*, 2014.

Resume

Cette thèse s'intéresse au développement de lois de commande basées sur la perception auditive. Dans le domaine de l'audition robotique, le contrôle du robot à partir d'informations auditives est généralement basé sur des approches de localisation de source sonore. Cependant, la localisation de source en conditions réelles est une tâche complexe à résoudre. En environnement intérieur, les perturbations causées par le bruit, la réverbération ou même la structure du robot peuvent altérer le processus de localisation. Cette tâche de localisation devient encore plus complexe si la source et/ou le robot sont en mouvement. Aujourd'hui, en se restreignant aux systèmes binauraux, la localisation sonore en environnement réel n'est pas encore réalisable de manière robuste. A l'opposé, nous développons dans cette thèse une commande référencée capteurs, l'*asservissement sonore*, qui ne nécessite pas de localiser la source. Le mouvement du robot est directement relié à la perception auditive: une tâche de positionnement est réalisée par une boucle de commande, où le mouvement du robot est régi par la dynamique d'indices sonores de bas niveau. Les résultats expérimentaux dans différentes conditions acoustiques et sur différentes plates-formes robotiques confirment la pertinence de cette approche en condition réelle.

Mots-clés: Robotique, Traitement du son, Asservissement sonore, Commande référencée capteur, Audition robotique, Différences interaurales.

Abstract

This thesis concerns the development of a control framework based on auditory perception. In general, in robot audition, the motion control of a robot using hearing sense is based on sound source localization approaches. However, sound source localization under realistic conditions is a significant challenge to solve. In indoor environment perturbations caused by noise, reverberation or even the structure of the robot may alter the localization process. When considering dynamic scenes where the robot and/or the sound source might move, the degree of complexity of source localization raises to a higher level. As a result, sound source localization considering binaural setup is not achievable yet in real-world environments. By contrast, we develop in this thesis a sensor-based control approach, *aural servo*, that does not require to localize the source. The motion of the robot is straightly connected to the auditory perception: a positioning task is performed through a feedback loop where the motion of the robot is governed by the dynamics of low-level auditory features. Experimental results in various acoustic conditions and robotic platforms confirm the relevance of this approach for real-world environments.

Keywords: Robotics, Audio signal processing, Aural servo, Sensor-based control, Robot audition, Interaural differences.