International Journal of Advanced Robotic Systems

OPEN ACCESS ARTICLE

Integration of the Multi-DOA Estimation Functionality to Human-Robot Interaction

Regular Paper

Caleb Rascon[1*], Ivan Meza[1], Gibran Fuentes[1], Lisset Salinas[1] and Luis A. Pineda[1]

1 Instituto de Investigaciones en Matematicas Aplicadas y en Sistemas, Universidad Nacional Autonoma de Mexico, Mexico
* Corresponding author(s) E-mail: caleb.rascon@iimas.unam.mx

Received 18 August 2013; Accepted 27 November 2014

DOI: 10.5772/59993

Abstract

Sound source localization is important in human interaction, such as in locating the origin of long-distance calls or facing other humans while in a conversation. It is of interest to apply such functionality to the core of human-robot interaction (HRI) and investigate its benefits, if any. In this paper, we propose three strategies for how to integrate the functionality of multiple directions-of-arrival (multi-DOA) estimation with a common scenario, in which the robot acts as a waiter while applying audio source localization. The proposed strategies are: a) the robot locates calls from users at a relatively long distance; b) the robot faces the user when taking the order; and c) the robot announces whether the acoustic environment is not conducive to understanding a speech command (mainly where more than one user speaks at once). It was seen that users react favourably to the functionality, and that it even has a noticeable influence on the success of the interaction.

Keywords DOA, HRI, Waiter, Multiple users

1. Introduction

Sound source localization plays an important part of human interaction, and thus it is of interest that it plays an important role in HRI. Although there have been several implementations of sound source localization in mobile robotic platforms (which will be described in more detail in Section 2), very few have delved into the specifics of how to integrate such functionality with the overall scheme of interaction between a human and a robot.

For this study, we will focus on the broadly-applied and essential HRI aspect of taking an order from different groups of users. There are many ways with which this class of information can be obtained, depending upon the current functionalities of the service robot. Face detection, speech recognition, gesture identification, etc., are all functionalities that can be used for this purpose. There are also functionalities that can complement the order-taking process, one of which is sound source localization.

In this paper, we explore three strategies, using the functionality of multi-DOA estimation to complement the process of obtaining an order from groups of users via Automatic Speech Recognition (ASR):

- *Contextualized long-distance calls*. The robot is able to estimate the origin of a call, for instance, "Waiter, waiter!" in a restaurant. For this, it uses its own location and the locations where users are most likely to be, together with estimations of the DOA of sound sources.
- *Facing the user*. The robot faces the user when he/she speaks. From the user's point of view, it seems as though

the robot is 'paying attention' to what he/she says, which complements the interaction.

- *Detecting multiple users*. The robot identifies multiple users speaking at the same time, and it lets them know that is not the best acoustic environment to recognize their orders.

These strategies, as well as their potential benefits, will be described in more detail in Section 3.

The rest of the paper is organized as follows: Section 2 gives a brief literature review of the use of multi-DOA estimation to HRI as well as a description of the challenges of implementation in a mobile robotic platform. Section 4 provides a review of the technical aspects of our in-house service robot, Golem-II+, and how multi-DOA estimation fits into its cognitive architecture, as well as a description of how it was implemented. Section 5 describes a case study of our robot as a waiter, in which the aforementioned order-taking aspects were the backbone of the interaction. Section 6 describes the evaluation of the integration of the proposed strategies in the case study, as well as presenting its results. Finally, in Sections 7 and 8, the results are discussed and conclusions are presented.

2. Background of multi-DOA estimation in HRI

Multi-DOA estimation is a basic component of sound-source localization, which is a vital ability in successfully interacting with the environment. Since it is omnidirectional and insensitive to occlusion and lighting conditions, auditory perception provides important complementary information to visual information for the identification and localization of interesting or potentially dangerous events in the environment. In fact, Perrot et al. [20] have pointed out that the information of the auditory spatial system serves as a guide to focus the attention of the visual system towards acoustic events which are outside the visual field. Humans in particular have a remarkable ability to localize sound sources, which helps them sense dangerous situations (e.g., identifying a car approaching) or perform social interactions (e.g., paying attention to a another person who is speaking).

2.1 Sound source localization applied in HRI

Sound source localization provides important information that allows face-to-face communication and proper interaction. Such behaviours in social interactions may include paying attention to a new sound source, moving towards it, or keeping facing a moving speaker.

Being aware of a speaker's location has allowed computer conversational systems to respond more naturally to users' needs. This has led to different face-to-face communication frameworks being proposed to enhance the interaction between humans and embodied agents [2, 5, 4, 18] and robots [32], although they have mostly been studied using 'Wizard of Oz' experiments. One example of a real application of these frameworks is the embodied conversational agent (ECA) of Bohus and Horvitz [3], which uses both visual and audio analysis to detect and track speakers in multi-party conversations. Another application of sound localization is to use the speaker's location for automatic camera steering, which is useful for teleconferencing and surveillance purposes [6].

The advantages of sound localization have inspired many robotics researchers to incorporate such an ability in robots so that they can better sense what is happening in their surroundings. For a robot, the direction of a sound can be a rich source of information that can help enhance its interaction with humans, for instance by letting users know that the robot is listening and waiting for orders, or simply showing the speakers that the robot is engaged in the conversation. Moreover, the direction of a human sound source can be further exploited by other modules in the robot, such as navigation, speech recognition and vision.

For example, an audio-visual speaker tracker was used to direct the attention of the robot SIG in multi-party interactions [19]. Similarly, Trifa et al. [34] have presented a system that integrates sound source localization with other functionalities for moving the robot's eyes and neck towards interesting events. Teachasrisaksakul et al. [31] developed an indoor robot navigation system based on sound localization. Recently, Li et al. [11] demonstrated a robot that can play the game *hide and seek*, in which it moves according to simple voice commands given by the players and localizes the player raising his/her hand after receiving the command "localization". Here, to localize the player, the robot makes use of sound localization and hand detection. More frequently, however, robot implementations of sound source localization have focused on enhancing speech recognition [39], separating multiple sound sources [36] or helping track objects [16] and persons [13].

Unfortunately, to the best of our knowledge, very little effort has been devoted to exploring how to integrate sound source localization systems in the overall scheme of HRI. This is surprising, as several of these systems have been integrated into several robots, as described above. One possibility for this is that on a service robot it is challenging to estimate the directions of multiple sound sources, and an unpredictable system may provoke user backlash, ruining the interaction. This is described in more detail in the following section.

2.2 Challenges in estimating multiple DOAs in a mobile robotic platform

As mentioned earlier, multi-DOA estimation is an essential part of sound source localization, which is a well-studied topic in signal processing. It has proven useful in applications ranging from fault monitoring in aircraft [30], to intricate robotic pets [10], to close-to-life insect emulation [9]. In addition, the principles employed in DOA estimation have been applied in the design of hearing aids [12].

A popular methodology for DOA estimation in robotic platforms is to use a microphone array with - usually - two microphones, as proposed in [17]. Adding more microphones generalizes the strategy, as a two-microphone array is an instantiation of classic reverse beamforming techniques [30], which create a noise map of the environment and then, by using metrics such as energy levels, propose possible sources of sound and their respective DOAs. However, to obtain a high resolution noise map, and thus a precise DOA estimation, beamforming techniques require a large quantity of microphones, which is impractical for mobile robotic platforms.

The topic of how many microphones to use in a service robot is intrinsic to the nature of the application, as it is important for the audio capture system to be mobile. A many-microphone solution can provide good results, such as the one proposed in [35] where the source signals were separated from each other in order to enhance speech recognition, and as a preamble for DOA estimation. However, it required an array of eight microphones positioned in a cube-like manner to work, doubling the height of the robot without it. On the other hand, a few-microphones solution, such as the two-microphone system presented in [15], may be light enough to be carried by a service robot, but might only work "adequately" (as stated by the authors themselves), presenting solutions in the limited -90°– 90° range.

A popular technique is the multiple signal classification (MUSIC) algorithm [28], which is able to detect the DOA of as many sources as one less the available microphones (e.g., one source with two microphones, two sources with three microphones, etc.). It does this by projecting the received signals in a DOA subspace based on their eigenvectors, similar to principal component analysis. It was applied in [14] with good results, although it has been observed that its performance decreases considerably in the presence of reverberation [38] (pp. 169).

Reverberation is one of the main challenges in the estimation of the DOA of a sound source, as it is prevalent in the locations where a service robot is expected to be, such as a restaurant. Moreover, it has been shown to hinder considerably the effectiveness of other current non-MUSIC-based DOA estimators [8].

Another challenge is that simultaneous speech is to be expected, as users may be speaking over each other. Moreover, no assumption can be made as to the location of the users relative to the robot, so their directions need to be estimated in the complete -179°– 180° range. Furthermore, because food/drink orders tend to be a few words-phrases, the whole process needs to be fast enough to carry out DOA estimations based on small-sized utterances of users.

As can be seen, carrying out multi-DOA estimation on a mobile robotic platform provides a unique challenge, as a balance needs to be struck. It needs to be light enough so as not to hinder the mobility of the service robot, meaning that the number of microphones needs to be maintained within a practical range in order to be carried by a service robot. At the same time, it needs to be able to handle a highly dynamic acoustic setting; thus, it needs to be robust enough on the software side to handle all of the aforementioned issues.

As mentioned before, this may be the reason why, with all the sound source localization systems implemented in service robots, very few have been formally integrated into an HRI scheme.

3. HRI-complementing strategies based on Multi-DOA estimation

There are several functionalities that can be derived from the information obtained via multi-DOA estimation. In this section, we present those that we propose for integration into an HRI scheme. As mentioned earlier, we are focusing on the HRI aspect of order-taking, which is observed in real-life scenarios. One of these scenarios, described in more detail in Section 6, is that of a waiter, which adds acoustic challenges while requiring the maintenance of user satisfaction.

The following strategies are proposed to complement order-taking in a HRI scheme.

3.1 Contextualized long-distance calls

The context of taking an order in an acoustically complex scenario implicitly specifies the location of the robot in its environment, as well as the possible places in which a user can be located (in the case of a restaurant, users could be at the bar, at other tables, etc.). If a user asks for the robot's attention, via the user's DOA, it is possible to make a viable estimation of the user's location. This provides for the possibility of knowing where to navigate next.

From the user's point of view, he/she is able to call the robot while navigating, even if the robot is not facing in the user's direction - although the robot may not be able to understand the actual command, the robot will be able to know where to go to retrieve it better. This presents a very natural way to communicate with the robot, similar to the way in which one might want to obtain the attention of a person from afar.

3.2 Facing the user

Having a robot rotate towards the user while speaking - and from the user's point of view - it appears as though the robot is facing him/her. This will provide the impression that the robot is 'paying attention' to the user when he/she is speaking to it, and may enhance the naturalness of the interaction. In the case of robots using directional microphones, this strategy will also provide the subtle cue that the robot will better understand a person who is in front of its microphone, and it may subtly indicate to the user how to better work with the robot.

3.3 Detecting multiple users

Current automatic speech recognizers are designed to recognize one speech command at a time. However, during a conversation between several users there can be situations when two or more speak at the same time. It is very useful, then, for a robot to be aware of these situations, since it will be an indicator that the speech recognizer will not perform well.

To integrate such behaviour in our robot, the following logic was utilized: For every speech command being recognized, the robot also checks for the number of DOAs detected during the utterance of the speech command. If only one DOA is detected, and an order is recognized, the robot can proceed with the following steps of the task. If there was more than one DOA detected, the robot should notify the users of the situation, and again carry out the order-taking process, though now coordinating with the users. This coordination can be carried out by, first, asking the users to speak one at a time, and then, going down the list of detected DOAs and, for each DOA in the list, facing that DOA and asking for an order in that direction.

4. Technical aspects of multi-DOA estimation on Golem II+

Golem-II+ is a service robot, presented in Figure 1, built with a primary focus of HRI. It is integrated by a cognitive architecture focused on HRI, termed an 'interaction-oriented cognitive architecture' (IOCA) [21], which can take advantage of different types of information interpreted from the environment, including the direction of the user. Because Golem-II+ is a conversational robot, it is of interest that it is able to detect and carry out conversations with several users at any point.

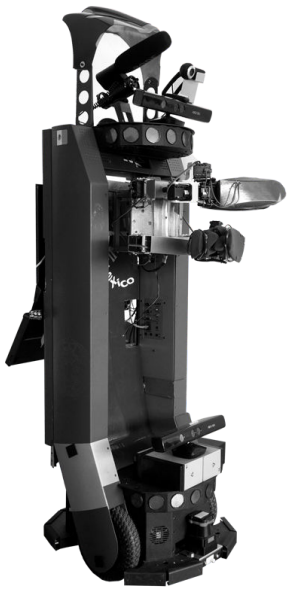


Figure 1. The Golem-II+ Service Robot

IOCA has a top level, called the 'Representation & Inference' phase, where the set of expected multi-modal situations are defined and ordered in what is called a 'dialogue model' (DM), which guides the HRI process and assumes an abstract-but-meaningful interpretation of the world. This interpretation is achieved by first obtaining an internal codification of the input [22] (i.e., the recognition phase), and then providing a meaning to that representation (i.e., the interpretation phase), guided by the contents of its memory and the current expectations of the DM. With this architecture, complex objectives can be fulfilled, such as a tour-guide robot capable of guiding a poster session [1].

A DM represents the protocol of an interaction. It can be seen as a graph composed of nodes (or *situations*) connected by arcs, which in turn are composed by pairs of *expectations* (α) and *actions* (β). Figure 2 illustrates the main elements of a DM.

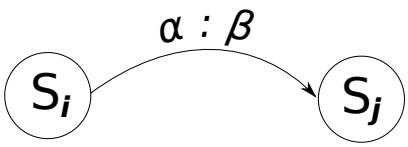


Figure 2. Graphical representation of a simple DM

An Expectation is met when the abstract interpretation of the world matches the expected occurrence in the world, such as the user greeting the robot at the start of a tour. If an expectation is met accordingly, the DM moves along the HRI process; if something happens in the world, but no expectations were in place for it in the current situation in the DM, the system executes a recovery DM to continue the HRI.

The multi-DOA estimation module resides inside the recognition phase, where every sound received is tagged with the characteristic of "direction". This characteristic is used by the interpretation phase to enrich the meaning of the information received by other modalities. A good example of this is when a user makes a food/drink order, which is being recognized by the ASR, and the DOA estimator complements this information by adding the direction in which the order was obtained.

The technique implemented in the innards of the multi-DOA estimation module is the lightweight multi-DOA estimator (LMDE) [27]. It was chosen because it provides a robust solution to the challenges described in Section 2.2, with a relatively lightweight hardware setup. In addition, there are aspects of this implementation that are taken advantage of in the HRI integration, which is particularly relevant to the case study. Thus, these technical aspects, as well as a complementary evaluation of their appropriateness to be applied in the explored case study, are described in detail in the following section. However, it is important to emphasize that the technique used may very well be any

other that is able to satisfy the requirements imposed by IOCA, as well as those discussed in Section 2.2.

4.1 Lightweight multi-DOA estimator

The LMDE [27] uses a triangular microphone array and is based on the parallel calculation of three inter-aural time differences (ITDs) for redundancy, which feeds a clustering-based DOA tracker. It comprises three modules:

1. *Audio Acquisition*. This obtains audio data from the microphones and provides it to the initial DOA estimation module.
2. *Initial DOA Estimation*. This estimates, from the audio data, an initial, fast but reliable DOA estimation of a single sound source in the environment.
3. *Multi-DOA Tracking*. This carries out dynamic clustering of the incoming DOA estimations, by which the DOAs of sound sources are proposed.

4.1.1 Audio acquisition

The audio acquisition in the LMDE requires that the audio from the three microphones be acquired simultaneously, in real-time. For this purpose, the JACK Audio Connection Toolkit [7] was employed for audio capture. It can sample at rates of 44.1 kHz and 48 kHz, providing a good resolution for ITD calculations without slowing down other robotic software modules.

4.1.2 Initial DOA estimation

The initial DOA estimation is carried out by the technique described in [26]. It avoids the problems that arise when estimating a DOA using 1D microphones arrays, and maintains a relatively light hardware setup: an equilateral-triangular array, positioned over the horizontal top panel of the robot, as shown in Figure 3 (which has a top view of the microphone array). To this effect, the system obtains a set of three simultaneous sample windows.

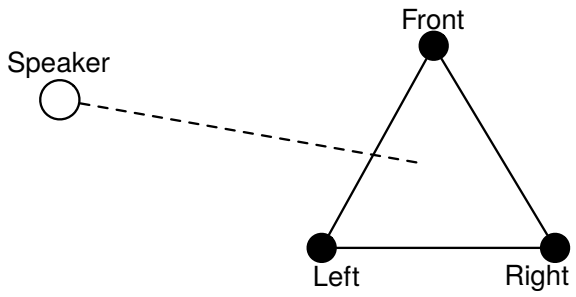


Figure 3. Hardware setup of the employed system

The audio data is passed through various serialized sub-modules: a band-pass filter, a voice activity detection (VAD) stage, multi-ITD estimation, a redundancy check and, finally, a final DOA estimation. The flow of data is summarized in Figure 4.

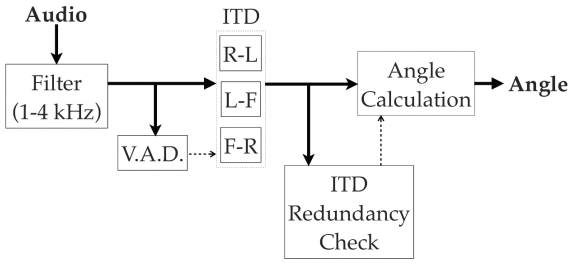


Figure 4. Initial DOA estimation flow of data

A general infinite impulse response band-pass filter is used at the beginning of the process to remove general ambient noise that is outside the human speech frequency bands. The filter model was created such that only frequencies between 1 and 4 kHz were let through. This reduces the sensitivity to unwanted noises that should always be ignored, such as high-pitch sounds, microphone hiss, etc. Concurrently, it did not degrade the sensitivity of the system in relation to human speech.

Next, VAD is carried out to trigger when to start and and when to stop DOA estimation. The VAD system adjusts the baseline of the environmental noise to any sound that is emitted with a pre-specified delay. In this way, the VAD system is able to work properly when the robot changes acoustic environments or when an unwanted noise becomes too loud.

When the VAD system is triggered, three possible ITDs are calculated using cross-correlation between sample windows R and L (I_{RL}), L and F (I_{LF}), and F and R (I_{FR}). An initial local DOA (D_{xy}^{m}) is calculated from each ITD (I_{xy}) using Equation (1):

$$D_{xy}^{m} = arcsin\left(\frac{I_{xy} \bullet V_{sound}}{F_{sample} \bullet d}\right) \tag{1}$$

where V_{sound} is the speed of sound (in *m*/*s*), F_{sample} is the sampling rate (in *Hz*), and d is the distance between microphones (in *m*).

Next, a pair of global DOAs (D_{xy}^{0} and D_{xy}^{1}) are calculated, using Equations (2) (3) (4). While the local DOA D_{xy}^{m} provides an angle from the perspective of the microphone pair alone, the global DOAs provide an angle from the perspective of the whole array. Specifically, D_{xy}^{0} is the global DOA calculated as if D_{xy}^{m} was coming from the front of the xy microphone pair, and D_{xy}^{1} is as if it was coming from the back:

$$\begin{aligned} D_{RL}^{0} &= \begin{cases} 0 & , if\ D_{RL}^{m} = 0 \\ -D_{RL}^{m} & , otherwise \end{cases} \\ D_{RL}^{1} &= \begin{cases} -180 + D_{RL}^{m} & , if\ D_{RL}^{0} > 0 \\ 180 + D_{RL}^{m} & , otherwise \end{cases} \end{aligned} \tag{2}$$

$$D^0_{LF} = \begin{cases} D^m_{LF} - 240 & , if\ (120 - D^m_{LF}) > 180 \\ 120 - D^m_{LF} & , otherwise \end{cases} \quad (3)$$
$$D^1_{LF} = D^m_{LF} - 60$$

$$D^0_{FR} = \begin{cases} D^m_{FR} + 240 & , if\ (-120 - D^m_{FR}) > 180 \\ -120 - D^m_{FR} & , otherwise \end{cases} \quad (4)$$
$$D^1_{FR} = D^m_{FR} + 60$$

These three global DOA pairs are used to check if the three ITDs are from a sound source located in the same angle sector. To do this, the average of the differences between the DOA pairs (C_{pqr}) is calculated using Equation (5):

$$C_{pqr} = \frac{|D^p_{RL} - D^q_{LF}| + |D^q_{LF} - D^r_{FR}| + |D^r_{FR} - D^p_{RL}|}{3} \quad (5)$$

where p, q and r, can be either 0 or 1. This provides eight possible C_{pqr}, the lowest of which is considered to be the *incoherence* of the sample window set. The global DOA set that is represented in the minimum C_{pqr}, meaning $[D^p_{RL},\ D^q_{LF},\ D^r_{FR}]$, is proposed as the DOA set for this window sample set.

Figure 5 shows two examples of two different datasets. Figure 5a shows a highly incoherent set, and it can be seen that no $[D^p_{RL},\ D^q_{LF},\ D^r_{FR}]$ combination "points" to a clear direction. On the other hand, Figure 5b shows a dataset configuration with very low incoherence, which is $[D^0_{RL},\ D^1_{LF},\ D^0_{FR}]$, with the combination p=0, q=1, r=0 which points in the direction of the speaker.

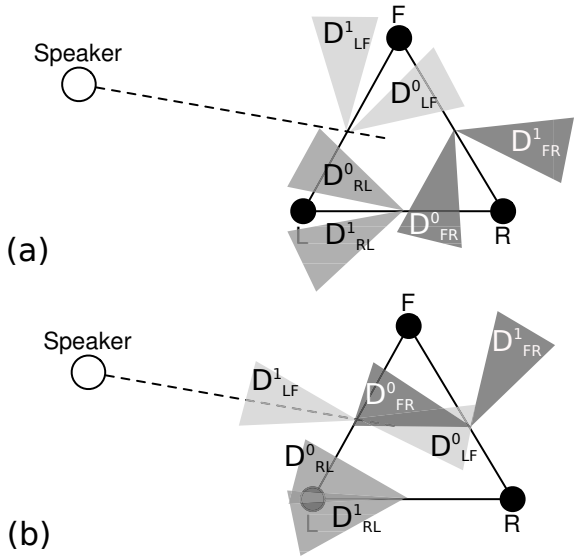


Figure 5. Example of window sets with (a) high incoherence and (b) low incoherence

A pre-specified *incoherence threshold* (measured in degrees of separation between the DOAs) is used to reject sample window sets. A high incoherence implies that the sample window set either has too much reverberation to be trustworthy for further processing, or that it contains **more than one sound source**. This rejection step serves as a type of redundancy check *per sampling window set*.

If the sample window set is considered coherent, its final DOA value (θ) is chosen from one member of its DOA set $[D^p_{RL},\ D^q_{LF},\ D^r_{FR}]$. The one that is chosen is the one that the ITD from which it was calculated has the lowest absolute value of the three (I_{RL}, I_{LF}, I_{FR}).

This final decision ensures that θ is based upon the microphone pair that is most perpendicular to the source and, because of the equilateral nature of the triangular array, this implies that it is always estimated using a local DOA D^m_{xy} with a value inside the -30°– 30° range (well within a close-to-linear area of Equation (1)). This means that all through the -179°– 180° range, there is always a close-to-linear ITD-DOA relation.

4.1.3 Multi-DOA tracking

The DOA estimator described in the previous section only provides results when there is considerable confidence in only one sound source being detected in a small sample window (up to 100 ms). It has been seen that, even in simultaneous-speech, users are not expected to talk with 100% overlap over each other. In fact, when analysing speech recognition, 'spurts' of non-overlapping speech have been considered to the order of 500 ms [29]. For example, in Figure 6, it can be seen how two randomly chosen tracks from the DIMEX corpus [24], when overlapping one another, still have some portions with no overlap between them.

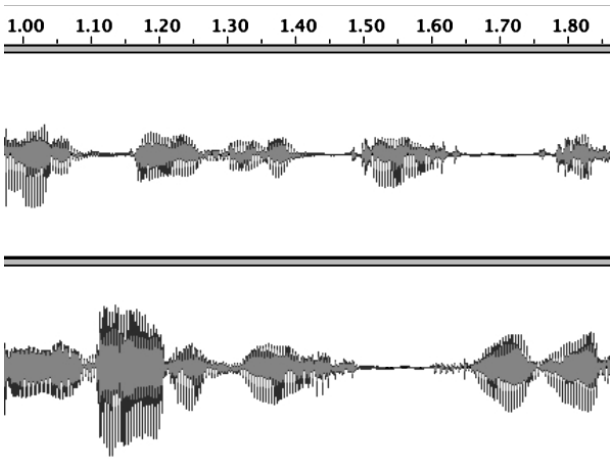


Figure 6. Non-overlapping simultaneous speech

This means that the initial DOA estimator is able to provide reliable results of single sources even in multi-user scenarios. However, because of the stochastic nature of the

presence of single-user sample windows in the simultaneous audio time line, such results would be provided in a sporadic fashion. To this end, a simple tracking system is employed that dynamically clusters similar DOAs into candidate sound sources.

The tracker maintains in its memory the last DOAs provided by the initial DOA estimator in a specific time frame. When a new DOA is estimated, the tracker carries out the following:

1. If the new DOA is not 'close enough' to the average DOA of any current cluster, or there are no clusters in the environment, create a new cluster with the new DOA.
2. If it is close enough to a current cluster, add the new DOA to it and re-calculate its new average DOA.

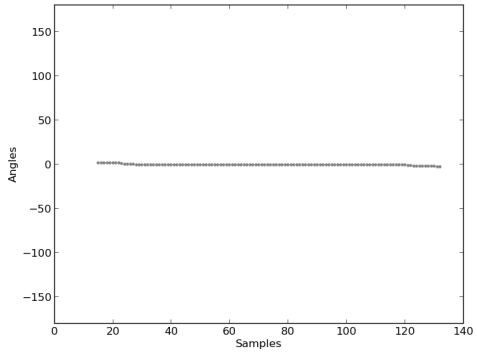
If a DOA is deemed to be too old, it is 'forgotten' by removing it from its cluster and re-calculating its average DOA.

Every cluster is considered a *candidate sound source* until it has more than a pre-specified number of DOAs attributed to it. At this time, it becomes a 'sound source' and its average DOA becomes its main estimated DOA.

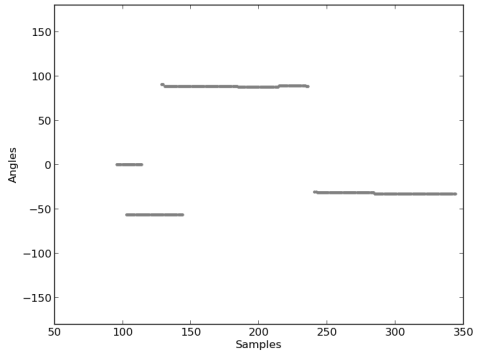
4.2 Complementary multi-DOA estimation evaluation

In Section 5, a case study is investigated that takes place in a typical indoor-social acoustic environment (medium reverberation, fan-noise, some chatter, etc.), with up to three users talking at the same time. It is of interest to know how adequate LMDE is in its application in such a case study. To this effect, a complementary evaluation was carried out that involved three tests, carried in a very similar setting and with the same acoustic environment to the one used in the case study: an indoor-social environment. The evaluation used one-to-three studio-grade monitor speakers simultaneously reproducing random recordings from the DIMEx100 Corpus [25] for 10 seconds, each monitor speaker acting as a sound source. For each number of sources, 10 tests were carried out. Figure 7 presents a representative output of the tests carried out with one, two and three simultaneous sources, respectively.

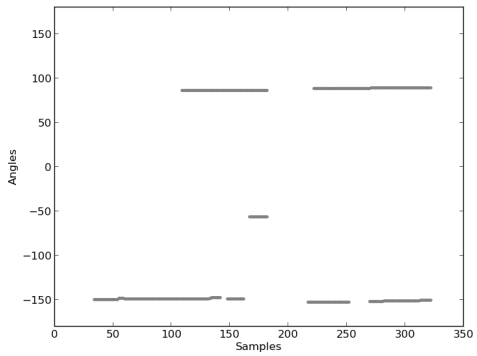
In Table 1, the results are shown in terms of precision, recall and F1 metrics, where an estimation was considered to be a true positive if the DOA of a source was reported and correctly estimated (with an error below ±10°) over a considerable amount of time (more than five seconds) during the 10-second lapse of each test. If a source that was present during the test was not reported over a considerable amount of time, it was considered to be a false negative. If a source that was not present during the test was reported over a considerable amount of time, it was considered to be a false positive.



(a) Multi-DOA output with 1 source (0°).



(b) Multi-DOA output with 2 sources (-30°, 90°).



(c) Multi-DOA output with 3 sources (-30°, 90°, -150°).

Figure 7. Tests with varying numbers of sources

Sources	Precision	Recall	F1
1	62.50	100	76.92
2	84.62	55.00	66.67
3	90.91	33.33	48.78
Average	*79.34*	*62.78*	*64.12*

Table 1. Results of an overall evaluation of the multi-DOA system

In Table 2, the results are shown again in terms of precision, recall and F1 metrics, but by observing the outputs far more strictly in a window-by-window manner, and by consider-

ing the correct estimation of the number of sources and their DOA (as well as the report of true positives, false positives and false negatives) for each sample window of the collected data.

Sources	Precision	Recall	F1
1	51.41	85.16	64.11
2	41.56	32.13	36.24
3	46.60	18.69	26.67
Average	*46.52*	*45.33*	*42.34*

Table 2. Results of a window-by-window evaluation of the multi-DOA system

These results show that there is definitely room for improvement for the LMDE while also showing its adequacy for use in the case study. In the overall evaluation, on average, the LMDE performed over 60%, which is very good considering the acoustic complexity of the setting. The window-by-window evaluation did not present as good results, although an average performance within the 42% range with such strict observation can be considered adequate. It is important to note that the objective of this evaluation was to see whether the LMDE is adequate to be employed in the case study, which it has shown that it is. A thorough evaluation of the LMDE in several acoustic scenarios and its improvement are outside the scope of this paper, but it is definitely a cause for future work.

5. Case study: Golem-II+ as a waiter

Golem-II+, carrying out the tasks of a waiter, was investigated as a case study to observe how the strategies described in Section 3 are carried out in a live setting, interacting with humans.

Such a task provides a good baseline to evaluate the service robot as a whole and the effect of the multi-DOA estimation specifically, as there are many elements involved that not only need to work well but need to work well *together*. Examples of these elements include navigation in a dynamic environment, face and gesture identification, object manipulation, speech recognition, all of which are of interest in the service robotics community. Additionally, this task involves the three behaviours we are interested in: a waiter can be called from a distance to a table; while taking orders, a waiter should face the user in turn to acknowledge him/her; and, if more than one user speaks at the same time, a waiter should direct how the ordering takes place.

The general concept of this task is that the users are customers in a restaurant who want the robot, acting as the waiter, to bring them drinks/food. There are several tables in the restaurant among which several users are seated. The location of the tables are known *a priori* to the robot[1]. When the robot is not doing anything, it is positioned in a place that is visually accessible to the users near the tables.

The users are able to call upon the robot, at any time, via voice. When a user does so, Golem-II+ uses its multi-DOA estimation module and the pre-known table locations to estimate from which table the call came from, and navigates towards it. Once it is at the table, it will ask for an order, following the logic of the *detecting multiple users* strategy: if more than one user talks at the same time, it will ask them to talk one at the time, and coordinate the order. During the coordination, Golem-II+ will face the user who is speaking.

Once Golem-II+ has the order of all the users at that table, it will go to the restaurant's bar and ask for them in pairs, if necessary (since the robot only carries one object at a time in each of its arms). At this stage, it can look for them over the bar using its vision system or ask directly for them.

As mentioned earlier, throughout the task, users at other tables which are not being attended can call for the attention of the robot. If the robot is busy, it will acknowledge the calling users by facing them and asking them to wait. For every call, Golem-II+ stores where the call came from so that it can later visit it.

5.1 Dialogue model

We modelled the task on the SitLog robotic language [23]. For this, we split the task into five main subtask models: *wait for call*, *walk and listen*, *ask orders*, *deliver order* and *arms*. In *wait for call*, Golem-II+ introduces itself and tells the user that it is ready to serve the table; in *walk and listen*, the robot navigates while paying attention to calls from the rest of the tables; in *ask order*, Golem-II+ takes orders while facing users or checking that only one speaks at a time; in *deliver orders*, the robot navigates to the 'bar' to retrieve the objects that the users asked for and delivers them appropriately; finally, in *arms*, Golem-II+ keeps track of what it has in each hand.

We follow an ISU approach to administer the next main action of the robot [33]. It employs two main variables: a list of tables from which users have called it and which need to be attended, and a list of orders taken per table that have not been delivered. Depending upon the state of these variables, it can choose its next action. Given the flexibility of the SitLog programming language, it is trivial to switch between two priorities: take orders first or deliver orders first.

To incorporate the multi-DOA estimation module into the DMs, we created a new type of expectation, *dirs*(*As*), which consists of a list of recently estimated DOAs.

Figure 8 shows part of the *walk and listen* sub-DM. While the robot is moving towards its destination *D* (triggered by the action *goTo*(*D*)), there are two possible situations that can interrupt its navigation: 1) that it has arrived at its

1 Waiters are expected to know this information, so this is consistent to the real-life scenario.

destination (expectation *arrived*), in which case the sub-dialogue ends, and 2) that the multi-DOA estimation module has detected a DOA (expectation *dirs*(*As*)) that is consistent with the direction of a table[2], in which case it turns towards the table, says "I will be there", registers the table for future reference, and resumes its navigation towards *D*. In the latter case, if the expectation *dirs*(*As*) has more than one consistent DOA, it turns towards each of them and registers them all.

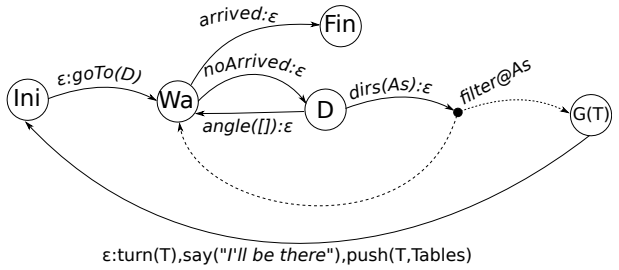


Figure 8. Dialogue model for the *walking and listening* sub-task

Figure 9 shows a part of the *ask order* sub-DM, which also relies on the multi-DOA estimation module. This model describes how Golem-II+ asks for an order, listens for an answer, and verifies that only one person talked. A DOA list is given to this model as an argument (P_s), from which its first item is removed, stored in *A* (action *pull* (*Ps, A*)), and is the direction in which the robot's neck turns to (action *turn*(*A*)); if *Ps* is empty, the robot continues in its default state (turned to front). Next, it asks the user for an order (action *ask_order*), which is followed by the act of listening for the order (expectation *order*(*X*)). Once the order has been provided, the robot consults the multi-DOA estimation module if any DOAs were detected during the act of listening (expectation *dirs*(*As*)), filtering them for consistency[3]. Depending upon the number of consistent DOAs detected, one of the following situations may be triggered: a) if no consistent DOAs were detected (situation *A*([])), it accepts the order and asks whether the user wants something else, but it does not face the user; b) if only one consistent DOA was detected (situation *A*([*A*])), it accepts the order, faces the user and asks the user whether he/she wants something else; or c) if more than one consistent DOA was detected (situation *G*(*As*)), it rejects the order, adds the DOAs to *Ps* (action *push*([*A,B,...*], *Ps*)), tells the users to speak one at a time, and returns to the initial situation to retake the order while providing *Ps* as an argument, which results in the robot facing each consistent DOA and taking an order for each one.

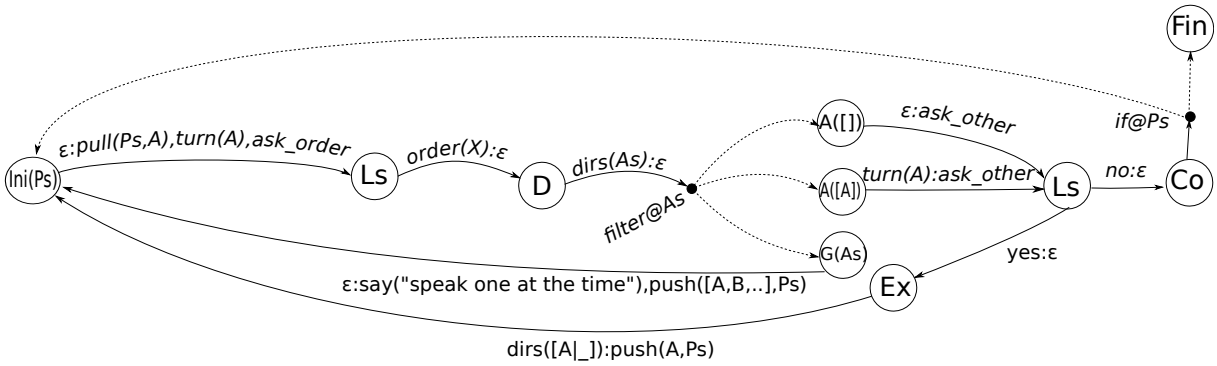


Figure 9. Dialogue model for the *ask order* sub-task

It is important to mention that the multi-DOA estimation technique used by the LMDE module can estimate DOAs with two different granularities: one provided by the initial detection phase, and the other by the tracking phase. Each type is used in different situations for the task.

For instance, when being called from a table (expectation *dir*(*As*) in the *walk and listen* mode), the estimated DOA value is only used as part of the consistency filter, and is not used directly to turn the robot towards the table. In this case, the DOAs provided by initial detection are more appropriate since they are obtained with less information but are more prone to errors.

In the case of order-taking (expectation *dir*(*As*) in the *ask order* mode), the estimated DOA value is used directly to face each user. In this case, the DOAs estimated in the tracking phase are more appropriate since these are considerably less error-prone (though they require more information to be provided).

2 Inconsistent DOAs that are not pointing to the position of a table are filtered-out by the function filter@As.

3 Inconsistent DOAs that do not point to possible positions of users sitting at the current table are filtered out by the function filter@As.

The two described sub-DMs are the core of the integration of the multi-DOA estimation module in our HRI scheme. A video with a demonstration of the task is available online[4].

6. Evaluation

To evaluate these strategies and how well-received they are by users, two forms of evaluations were carried out, and are detailed as follows. One is by an evaluation carried out by volunteers that participated as customers in the waiter scenario. The second is via presenting this scenario with expert users as a demonstration in the RoboCup@Home 2013 international Service Robotics competition.

6.1 Evaluation via volunteers as customers

In Figure 10, a description of the evaluation scenario is presented. As can be seen, three users were positioned in two tables, two of them shared a table and one was by himself/herself. There was also a bar, from which drinks could be picked up.

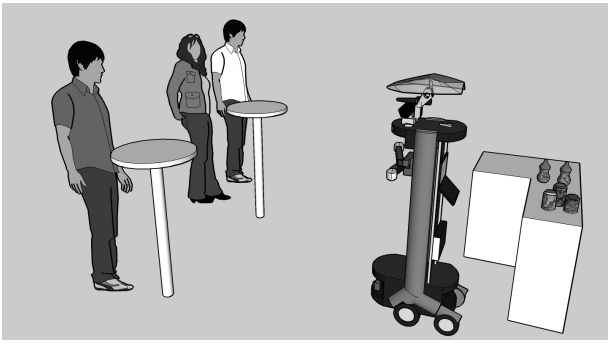


Figure 10. Waiter Scenario

We ran the task and let the users compete for the attention of the robot. Since order-taking was the main focal point of evaluation, we made the robot prioritize taking orders (instead of delivering them), and we did not allow the delivery of the orders. We instructed the users to try to get their order as quickly as possible and measured the success of the interaction for each of them in terms of being able to make their order.

In order to obtain a good indicator of the performance of our system, we let users interact with the robot only once: there were only vague descriptions of how the interaction would play out, as the users were only told that the role of Golem-II+ was that of a waiter and that it would only use its ears to locate them. The intent behind this was to let the users figure out how to interact with the robot as soon as the interaction began. We only gave the robot one chance to take an order from a user. When the robot started to deliver the order, the interaction was stopped. If a user was not able to get an order in, it was considered to be an unsuccessful interaction for that user.

Table 3 summarizes some of the characteristics of this evaluation. There were 30 volunteers who participated in 10 interactions, normally lasting 10 minutes. On average, the volunteers were 21 years old, the youngest being 19 and the oldest 28 years old. The possible orders that the users could give consisted of two Mexican traditional drinks. 60% of the users did not have any previous experience of interacting with a robot, and none of them had previous experience with this task.

Our evaluation was carried out in Spanish, since it is the native language of the users. However, we have an alternate version in English.

Interactions	10
Users	30
Language	Spanish
Gender	Females: 16.7%, Males: 83.3%
Average age	23
Experience	Golem: 36.7%, other robot: 3.3%
	none: 60.0%

Table 3. Summary of characteristics of the interactions of the evaluation

After each interaction, we collected information about the satisfaction of the users by means of a questionnaire based on the *Paradise* framework [37]. In order to account for the interaction through DOAs, we added the *Did the robot hear you?* question to that framework, as well as a yes/no question: *Did you like the rhythm?*. The resulting questionnaire consisted of eight questions with a Likert scale of 4 options, from positive to negative. Table 4 summarizes the eight measured aspects within the framework. In addition to the questionnaire, we also collected performance information by means of the system logs and our own notes.

The results are normalized with the weights 1, 0.66, 0.33 and 0, from the most positive to the least positive.

Voice intelligibility	81.1%
Understanding by ASR	57.8%
Multi-DOA Performance	58.9%
Figuring out interaction	72.2%
Speed	55.6%
Expectations	70.0%
Future use	64.4%
Stress	80.0%

Table 4. Summary of the results of the opinions of users interacting with Golem-II+ as a waiter

4 http://golem.iimas.unam.mx/waiter

As can be seen in Table 4, the opinions of the users about the robot are, in general, positive. The expectation and stress level were regarded very positively by the users. Moreover, considering that the evaluation setting was a very strict one, the fact that the future use aspect is above 50% can be considered to be good. These three aspects are of great importance from the HRI standpoint, as it implies that the users were comfortable during the interaction, that the robot reacted in ways that they expected, and that the users were open to using it again in the future.

From Table 5, we can gather that the multi-DOA estimation module was used by the users in all of the interactions. In addition, the robot was almost always able to return feedback to the users calling it. Another interesting point is that the detection of multiple users speaking at the same time was used in 40% of the interactions. Considering that the situation did not account for people talking over each other, this equates to a significant amount of use, heavily implying that multiple source detection is of significant importance in an order-taking interaction.

Multi-DOA activated	100.0%
Calling	100.0%
Robot reacted	96.7%
Spoken at the same time	40.0%
Ordered	60.0%
Not ordered	40.0%
Interactions with one order	90.0%
Interactions with two orders	50.0%
Interactions with three orders	40.0%

Table 5. Summary of the events during the interactions

However, user opinion is not as positive about the robot's comprehension skills for both the speech recognizer and the multi-DOA estimation module. We believe that this is partly because 40% of the users were not able to order. However, in 40% on the interactions, *all* the users were able to order. In addition, in 90% of the interactions, there was at least one ordering, and only in 1 interaction was nothing ordered.

6.1.1 Discussion on interaction success vs. perception of being listened to

An additional result is the relationship that was found between the success of the interaction and the users' perceptions of being listened to. As described before, the success of an interaction is measured per-user, depending upon whether he/she was able to provide an order to the robot. In Figure 11, these aspects are plotted against each other.

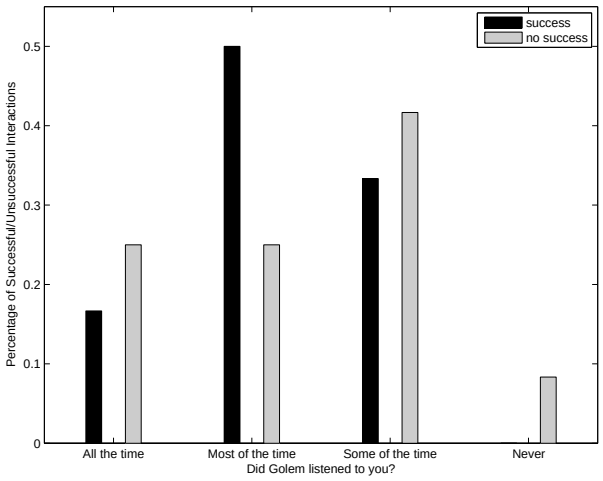


Figure 11. Percentage of successful and unsuccessful interactions vs. the perception of the user being listened to

As can be seen in Figure 11, there is a vague overall tendency of successful interactions between the robot and users who perceived the robot to be a good listener, as well as a tendency of unsuccessful interactions between the robot and users who perceived the opposite. This tendency is somewhat representative, as nearly half of the users (40%) were able to successfully provide an order, while the other near-half (60%) were not.

However, it is important to note that, first, this result was observed from data that was not aiming for its evaluation; second, no statistical significance was found in this result; and third, this tendency is not consistent throughout all the data (the information provided in the bin called "All the time" in Figure 11 breaks this tendency). All of these factors imply that this result should by no means be considered definitive, and it definitely calls for more subsequent studies. Still, the fact that it appeared provides encouragement in that robust sound source localization functionality may in some way influence the success of a HRI.

6.2 Evaluation via demonstrations in the service robotics competition

The alternate English version of the waiter case study was presented as a demonstration in the *@Home* league of the RoboCup 2013 international competition, which took place in Eindhoven, The Netherlands. It was carried out as part of the Open Challenge test, which requires the robot to demonstrate an application in relation to an open-ended topic. It is evaluated in terms of demonstration success and fluidity of interaction by a panel of judges composed of representatives of all the participating teams. Golem-II+ was given the top score among all the teams.

In addition, the Technical Committee and the Executive Committee of the competition gave Golem-II+ the 2013 Innovation Award for this demonstration. This award is provided only to robots that present novel concepts that are carried out successfully in conditions of competition.

7. Discussion on evaluations results

All the evaluations - taken together - show that the strategies proposed to integrate a multi-DOA estimation module in a HRI scheme were generally well-received by both novice and expert users, and that they are in constant use during the interaction.

The volunteers-as-customers evaluation in itself also showed that the concept of a DOA is quickly and appropriately adopted by novice users as part of their interaction with the robot. The evaluation by demonstrations-in-service-robotics-competition itself also showed that expert users welcome these strategies for its success in complex acoustic settings, and that they see the potential for use in their already-established interaction protocols.

All of these results imply that the strategies are a good step in the integration of a multi-DOA estimation module in a HRI scheme, but also that there is definite work to carry out to complete this integration in both HRI strategies as well as in the technical aspects of multi-DOA estimation.

8. Conclusion

Multi-DOA estimation, although challenging when carried out in a mobile robotic platform, provides acoustic information that can strongly complement HRI, specifically order-taking. Three strategies involving multi-DOA estimation were proposed:

1. Having specific areas where the user can be located in the environment, a DOA value can be used to estimate the location of the user when he/she speaks to the robot from afar. The robot does not need to be facing the user to make this estimation, and the user can call out and interrupt the robot at any moment.
2. Facing the user who is speaking provides a natural response from the robot, as though it is 'paying attention' to the user, complementing the interaction.
3. Detecting various users speaking at the same time provides the robot with the ability to not only ascertain whether the acoustic environment is conducive for good speech recognition, but also to coordinate with users as to who goes first in providing a command.

As a case study, these strategies were observed and verified by implementing them in our service robot, Golem-II+, as part of the task of being a waiter in a restaurant. This task was carried out by several non-experienced volunteers and, in general, they reported that they were comfortable with the robot's behaviour, that the robot reacted in ways that they expected, and that they would be open to using the robot again in the future. As a confirmation of these findings, the English version of this task received the Innovation Award from the *@Home* league of the RoboCup 2013 competition.

As future work, it would be of interest to improve the multi-DOA estimation system, as it is definitely called for, according to both evaluations carried out (global and window-by-window). Moreover, a vague tendency was observed between the success of an interaction and the user's perception of being listened to, which calls for further study. Furthermore, all three strategies were evaluated as a "group", since the case study asked for their intertwined application; an individual evaluation of each strategy would have been unreasonable, since one relied on the other to carry out the task, and finding a proper substitute for each strategy was beyond the scope of this work. However, it would be possible, with another case study, to investigate the impact of the proposed strategies on their own.

In addition, we propose adding and investigating the benefits of a fourth strategy: *automatic user labelling*. Instead of embarking on a user identification process for every order taken (such as asking for a name or searching for faces in the visual range), the DOA estimated while listening for an order could be used to automatically label the user. From the user's point of view, this would make the interaction more efficient and closer to a natural interaction during the provision an order.

Moreover, it is of interest to combine the multi-DOA estimation module with visual information for redundancy purposes, such as in the case where users switch places at the table while the robot is retrieving their order.

9. Acknowledgements

The authors thank the support of CONACYT through projects 81965 and 178673, PAPIIT-UNAM through project IN115710-3, and ICYTDF through project 209/12.

10. References

[1] Héctor Avilés, Montserrat Alvarado-Gonzalez, Esther Venegas, Caleb Rascón, Iván V. Meza, and Luis A. Pineda. Development of a Tour-Guide Robot Using Dialogue Models and a Cognitive Architecture. *Advances in Artificial Intelligence - IBERAMIA 2010*, 6433:512–521, 2010.

[2] Dan Bohus and Eric Horvitz. Computational Models for Multiparty Turn Taking. Technical Report MSR-TR 2010-115, Microsoft Research, 2010.

[3] Dan Bohus and Eric Horvitz. Facilitating multiparty dialog with gaze, gesture, and speech. In *Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 5:1–5:8, 2010.

[4] Dan Bohus and Eric Horvitz. Decisions about turns in multiparty conversation: from perception to action. In *Proceedings of the International Conference on Multimodal Interfaces*, pages 153–160, 2011.

[5] Dan Bohus and Eric Horvitz. Multiparty turn taking in situated dialog: study, lessons, and directions. In

Proceedings of the SIGDIAL Conference, pages 98–109, 2011.

[6] B. Dahlan, W. Mansoor, M. Abbasi, and P. Honarbakhsh. Sound source localization for automatic camera steering. In *Networked Computing and Advanced Information Management (NCM), 2011 7th International Conference on*, pages 20–25, June 2011.

[7] Paul Davis. JACK Connecting a World of Audio. http://jackaudio.org/.

[8] Christof Faller and Juha Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *The Journal of the Acoustical Society of America*, 116 (5): 3075–3089, 2004.

[9] A. D. Horchler, R. E. Reeve, B.H. Webb, and R. D. Quinn. Robot Phonotaxis in the Wild: a Biologically Inspired Approach to Outdoor Sound Localization. In *Sound Localization," 11 th International Conference on Advanced Robotics, (ICAR '03)*, pages 1749–1756, 2003.

[10] Rong Liu and Yongxuan Wang. Azimuthal source localization using interaural coherence in a robotic dog: modeling and application. *Robotica*, First View: 1–8, 2010.

[11] Xiaofei Li, Miao Shen, Wenmin Wang, and Hong 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 (78), 2012.

[12] Michael E. Lockwood, Douglas L. Jones, Robert C. Bilger, Charissa R. Lansing, William D. O'Brien Jr., Bruce C. Wheeler, and Albert S. Feng. Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. *Journal of the Acoustical Society of America*, 115 (1): 379–391, January 2004.

[13] Ivan Markovic and Ivan Petrovic. Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering. *Robot. Auton. Syst.*, 58 (11):1185–1196, 2010.

[14] Satish Mohan, Michael E. Lockwood, Michael L. Kramer, and Douglas L. Jones. Localization of multiple acoustic sources with small arrays using a coherence test. *Journal of the Acoustical Society of America*, 123 (4):2136–2147, April 2008.

[15] John C. Murray, Harry R. Erwin, and Stefan Wermter. Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks. *Neural Networks*, 22 (2):173–189, 2009.

[16] Kazuhiro Nakadai, Ken ichi Hidai, Hiroshi Mizoguchi, Hiroshi G. Okuno, and Hiroaki Kitano. Real-time auditory and visual multiple-object tracking for humanoids. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 2, pages 1425–1432, 2001.

[17] Kazuhiro Nakadai, Hiroshi G. Okuno, and Hiroaki Kitano. Real-Time Sound Source Localization and Separation for Robot Audition. In *in Proceedings IEEE International Conference on Spoken Language Processing, 2002*, pages 193–196, 2002.

[18] Yukiko I. Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, volume 1, pages 553–561, 2003.

[19] Hiroshi G. Okuno, Kazuhiro Nakadai, and Hiroaki Kitano. Social Interaction of Humanoid Robot Based on Audio-Visual Tracking. *Developments in Applied Artificial Intelligence Lecture Notes in Computer Science Volume*, 2358:725–735, 2002.

[20] David R. Perrott, Kourosh Saberi, Kathleen Brown, and Thomas Z. Strybel. Auditory psychomotor coordination and visual search performance. *Perception and Psychophysics*, 48:214–226, 1990.

[21] Luis A. Pineda, Iván V. Meza add Héctor H. Avilés, Carlos Gershenson, Caleb Rascón, Montserrat Alvarado, and Lisset Salinas. IOCA: An Interaction-Oriented Cognitive Architecture. *Research in Computer Science*, 54:273–284, 2011.

[22] Luis A. Pineda, Ivan V. Meza, and Lisset Salinas. Dialogue Model Specification and Interpretation for Intelligent Multimodal HCI. *Advances in Artificial Intelligence – IBERAMIA 2010 Lecture Notes in Computer Science*, 6433:20–29, 2010.

[23] Luis A. Pineda, Iván V. Meza Lisset Salinas, Caleb Rascón, and Gibrán Fuentes. SitLog: A Programming Language for Service Robots' Tasks. *International Journal of Advanced Robotic Systems*, 2013. To appear.

[24] L. A. Pineda, H. Castellanos, J. Cuétara, L. Galescu, J. Juárez, J. Llisterri, P. Pérez, and L. Villaseñor. The Corpus DIMEx100: Transcription and Evaluation. *Language Resources and Evaluation*, 44:347–370, 2010.

[25] L. A. Pineda, H. Castellanos, J. Cuétara, L. Galescu, J. Juárez, J. Llisterri, P. Pérez, and L. Villaseñor. The Corpus DIMEx100: Transcription and Evaluation. *Language Resources and Evaluation*, 44:347–370, 2010.

[26] Caleb Rascón, Héctor Avilés, and Luis A. Pineda. Robotic Orientation towards Speaker for Human-Robot Interaction. *Advances in Artificial Intelligence - IBERAMIA 2010*, 6433:10–19, 2010.

[27] Caleb Rascon and Luis Pineda. Lightweight Multi-direction-of-arrival Estimation on a Mobile Robotic Platform. *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2012*, I:665–670, 2012.

[28] R. Schmidt. Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on*, 34 (3):276–280, 1986.

[29] Elizabeth Shriberg, Andreass Stolcke, and Don Baron. Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation. In *in Proceedings of Eurospeech 2001*, pages 1359–1362, 2001.

[30] M.G. Smith, K.B. Kim, and D.J. Thompson. Noise Source Identification Using Microphone Arrays. *Proceedings of the Institute of Acoustics*, 29 (5), January 2007.

[31] K. Teachasrisaksakul, N. Iemcha-od, S. Thiemjarus, and C. Polprasert. Speaker tracking module for indoor robot navigation. In *Proceedings of the International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology*, pages 1–4, 2012.

[32] Andrea Lockerd Thomaz and Crystal Chao. Turn-Taking Based on Information Flow for Fluent Human-Robot Interaction. *AI Magazine*, 32 (4):53–63, 2011.

[33] David Traum and Staffan Larsson. The Information State Approach to Dialogue Management. In Jan van Kuppevelt and Ronnie Smith, editors, *Current and New Directions in Discourse and Dialogue*, pages 325–353. 2003.

[34] V.M. Trifa, A. Koene, J. Moren, and G. Cheng. Real-time acoustic source localization in noisy environments for human-robot multimodal interaction. In *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication*, pages 393–398, 2007.

[35] JeanMarc Valin, Jean Rouat, and François Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In *in Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pages 2123–2128, 2004.

[36] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, and H.G. Okuno. Robust Recognition of Simultaneous Speech by a Mobile Robot. *IEEE Transactions on Robotics*, 23 (4):742–752, 2007.

[37] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics, 1997.

[38] DeLiang Wang and Guy J. Brown, editors. *Computational auditory scene analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley-Interscience, 2006.

[39] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H.G. Okuno. Improvement of robot audition by interfacing sound source separation and automatic speech recognition with Missing Feature Theory. In *Proceedings of the IEEE International Con*ference on Robotics and Automation, volume 2, pages 1517–1523, 2004.