Hindawi Publishing Corporation
Journal of Robotics
Volume 2016, Article ID 3081048, 13 pages
http://dx.doi.org/10.1155/2016/3081048

Research Article

On Indexicality, Direction of Arrival of Sound Sources, and Human-Robot Interaction

Ivan Meza, Caleb Rascon, Gibran Fuentes, and Luis A. Pineda

Instituto de Investigaciones en Matematicas Aplicadas y en Systemas, Universidad Nacional Autonoma de Mexico, Ciudad Universitaria, Coyocan, 04510 Mexico City, MEX, Mexico

Correspondence should be addressed to Ivan Meza; ivanvladimir@turing.iimas.unam.mx

Received 30 November 2015; Revised 7 March 2016; Accepted 10 April 2016

Academic Editor: Gordon R. Pennock

We present the use of direction of arrival (DOA) of sound sources as an index during the interaction between humans and service robots. These indices follow the notion defined by the theory of interpretation of signs by Peirce. This notion establishes a strong physical relation between signs (DOAs) and objects being signified in specific contexts. With this in mind, we have modeled the call at a distance to a robot as indexical in nature. These indices can be later interpreted as the position of the user and the user herself/himself. The relation between the call and the emitter is formalized in our framework of development of service robots based on the SitLog programming language. In particular, we create a set of behaviours based on direction of arrival information to be used in the programming of tasks for service robots. Based on these behaviours, we have implemented four tasks which heavily rely on them: following a person, taking attendance of a class, playing Marco-Polo, and acting as a waiter in a restaurant.

1. Introduction

Knowing the origin of a sound source is an important skill for a robot. Usually, this skill is associated with survival behaviours in which this knowledge is used to react to natural predators or eminent danger. However, this skill also plays a key role during interaction, for instance, in calling a waiter over. At first, knowing the direction of arrival (DOA) of a sound source may seem too basic to be an element in the interaction; however, as we will show, this information can be meaningful so that the robot can react in a proper manner and enhance the interaction.

We consider the direction of arrival information as an index in the sense proposed by Peirce's sign theory [1, 2]. In this theory, there are three types of signs given their relation with the represented object. These types are *icon*, *index,* and *symbol*. *Icons* reflect qualitative features of the object, for instance, pictures or drawings of the object. *Indices* have an existential or physical connection with the objects in a specific context; for instance, a dark cloud is a sign for rain. On the other hand, *symbols* have a stable interpretation based on a convention that connects them to the object, for instance, a logo of a product. This also holds for spoken communication; onomatopoeic words, for instance, are iconic since they resemble what they represent (e.g., *chop*), pronouns are indices since they "point" to the object they represent, and nouns are symbolic since they are conventional and detached from the object they represent. In the context of this classification, DOA information has an indexical quality; it is an index of the position of the object which emits the sound and also of the object itself, since it holds a strong connection between the emitter of a sound (object) and the direction of arrival (sign). In this work, we exploit the use of this index to support the interaction between a robot and the users.

The present work of indexical DOAs is formalized and implemented in the SitLog programming language that we have developed and used to program our service robot [3, 4]. SitLog defines a set of behaviours which range from simple (one skill) to composed (more than one skill) behaviours. These behaviours are the basic blocks used to program our robot at a higher level. Examples of simple behaviours are *walk*, which takes the robot from its current position to a destination, and *ask*, which prompts the user with a question and waits for the spoken answer. An example of a composed behaviour is *searching for an object* in different locations,

since it relies on other basic behaviours such as *walk* and *see object*. In essence, a behaviour has to be (1) portable, so it can be used in different tasks, (2) compositional, so by coupling it with different behaviours we can create more complex behaviours or program a task, and (3) able to handle potential failures (e.g., not arriving to its destination, not listening to an answer from the user, or not finding the object).

In order to model the indexicality of the DOAs and incorporate it into the interaction capabilities of our robot, we propose a simple behaviour to handle DOAs and interpret them as indices. This supports the interaction by allowing calls at a distance. In this case, the user or users can bring the attention of the robot to a specific area in which the sound originates. This behaviour can also be used in combination with the *walk* behaviour to allow interruptions during walking such as calling a waiter over. Additionally, the behaviour also supports the verification of the number of sources during a conversation. In particular, this can be used in combination with the *ask* behaviour to ensure that when the robot is listening to an answer, there is only one person talking.

This paper is organized as follows: Section 2 presents previous work in which sound information is part of a task for a service robot. Section 3 reviews indexicality as interpretation procedure. Section 4 presents the framework we use to interpret DOAs signs as indices. Section 5 describes the manner in which the DOA behaviour resolves a DOA index. Section 6 shows the four tasks on which the indices from sound sources are used to direct the interaction. Section 7 presents a discussion about our proposal and findings.

2. Previous Work

The use of sound source localization (SSL) in robotics is a blooming field; Okuno and Nakadai and Argentieri et al., present reviews of the main methods and their use with different types of robots [6–8]. Since its application to robotics, SSL has been promoted as a main skill for robots. Brooks et al. proposed it as a basic skill for interaction in the Cog project [9]. The robot SIG was proposed as an experimental setting for the RoboCup competition [10]; and the robot Spartacus participated in the 2005 AAAI Mobile Robot Challenge implementing SSL as a part of its skill set [11].

Although the interaction goal is a driving force in the field, there is a great effort on developing a robust SSL module. With this goal in mind, the preferred setting is the use of an array of microphones on board of the robot. A many-microphone solution provides good performance given its redundancy. In [12], Valin et al. propose an 8-microphone 3D array to enhance speech recognition and as a preamble for DOA estimation. More microphones are possible. Hara et al. use two arrays of 8 microphones (16 in total) [13]. However, minimal systems with 2 microphones are also possible, such as the one presented in [14]. A common setting, which we follow, is to use 3 microphones [15, 16]. The best approach and configuration are still an open question. Badali et al. present an evaluation of the main approaches over an 8-microphone array for mobile robots [17].

Source localization can be reactive as a modality of the interaction as presented in works that deal with following a conversation [18–21]; or it can be an essential part of the interaction, such as using sound localization to control the navigation of the robot. In particular, Teachasrisaksakul et al. propose a system that follows a person through a room by her or his voice [22]. On the other hand, Li et al. demonstrated a robot which plays *hide and seek*, using visual and audio modalities in the interaction [23]. These works follow a similar approach on how to incorporate sound source location as a part of the interaction as the one we propose. However, since only one user is present in their work, its treatment as an index is trivial. The DOA sign is directly interpreted as the user; there are no conflicts regarding who or what the sign could represent.

Given the progress in the field, there has been an effort directed at capturing a more complicated interaction. For instance, Fransen et al. present a multimodal robot which follows the instructions among two users to reach a target [24]. Nakadai et al. present a robot which is able to judge rock-paper-scissors sound games [25]. The Quizmaster robot plays a game in which it asks a question to up to four participants and uses SSL and source separation to decide who answered it first [26]. Do et al. present a robot which together with a caregiver log and detect the origin of certain sounds [27]. At this level of complexity of the interaction, DOA signs are being indirectly used as indices: the DOAs become indices which signify directly the users and this information is used to disambiguate who calls or wins. In [27], the DOAs and a category associated with them signify types of events. At this level, the multiple sources make a method to assign the DOA to the right entity necessary. As we will see in this work, the benefit of considering a DOA as an index in the right context allows us to use indexical resolution to perform this disambiguation in the cases that more than one DOA competes. In this work, we formalize this resolution similarly to other reference resolution mechanisms on robots, such as deictic references [28, 29]. This differs from other approaches in which DOAs and multimodal information complement each other for disambiguation [30].

3. Indexical Expressions

An interactive agent such as a robot has to understand the intention of its users or have expectations about the events that can occur in the environment in order to produce an adequate response. Considering the theory of signs by Pierce, we can formulate the goal of "understanding" as providing a sign or a set of signs to identify the signified object or objects [1, 2, 31]. In order to denote the object, the robot builds a representation of it. For example, if the user commands *robot go to the kitchen,* the sequence of symbols *robot*, *go*, *to*, *the,* and *kitchen* get translated into the representation *go*(*robot*, *kitchen*), with which the robot can establish the signified objects (itself, the kitchen, and the action) and act accordingly. A similar mechanism can be used for the interpretation of iconic signs; for instance, when the robot sees a table and identifies on it a bottle of juice, it will

Table 1: Examples of indexical expressions and their resolution.

Sign	Representation	Constraint	Resolved
This is juice + pointing gesture	$\lambda x.name(x : pose, juice)$	Pose of object	$name(object, juice)$
Take it	$\lambda x.take(x : object)$	Object	$take(juice)$

generate the representation $on(juice, table)$, which denotes the corresponding juice object fully.

When the set of signs of the interaction includes indexical signs, the representation is not fully specified by the signs since the index has to be linked in an extra resolution step to the denoted object. For instance, if the user states the command *robot come here*, the representation for this utterance is underspecified as $\lambda h.go(robot, h)$. In order to fully specify this expression, it has to be analysed in relation to the context of the task, in which the position of the user can be inferred. For example, if the user is located in the kitchen, the expression should be resolved to $go(robot, kitchen)$.

In this scheme, indexical signs produce an underspecified representation at first which should be later resolved in the presence of contextual information. Furthermore, indices provide additional constraints on the type of object which can resolve the expression; for instance, in the previous example, the destination of the robot has to be resolved into a valid position. In order to account for this phenomenon, inspired by the treatment of indices in [32], we propose the following formulation for an index in lambda calculus in which we use the symbol ":" to signal a constraint α over a certain variable x:

$$\lambda \alpha Q x.\,(Q@x : \alpha)\,. \tag{1}$$

Indices defined in this form receive as parameters the constraint α and the underspecified predicate Q; the effect of applying such arguments is the binding of the constrained variable $x : \alpha$ into the expression Q. In this case, the symbol @ signifies a functional application (i.e., β reduction in lambda calculus.) If we apply the index in our previous example $Q = \lambda h.go(robot, h)$ and $\alpha = user_{pos}$, the following representation is produced: $\lambda h.go(robot, h : user_{pos})$. It is still underspecified, but it has constrained the variable h to be the position of the user. A second process of resolution has to identify from the context the fact that the user is in the kitchen.

Table 1 shows two examples of indexical signs using this formulation with a robot. For these examples, consider the following context: the user and the robot are in front of a table with an orange juice placed on it. The first illustrated example is a multimodal expression: the user says something and points to an object. The word *this* in conjunction with the pointing gesture define an index which imposes a spatial constraint on the object being signified by both. In this case, the constraint is the pose of the object (position and orientation extracted from the visual system). This constraint is used to identify the object in the scene, which in the example resolves into the bottle of juice. The second example shows an indexical reference with the word *it* which also adds a constraint [33]. In this case, the underspecified variable has to be resolved to an object in the real world, the bottle of juice again. Notice that we have invoked a spatial and deictic resolution; however, we have not specified a particular mechanism to perform this type of resolutions. The topic of effective mechanisms for the resolution of indexical expressions continues to be studied in different fields [34–36], and which mechanism is applied depends on the type of expression and the type of task being solved. In this work, we define the mechanism for the resolution of indexical DOA expressions and the conditions in which it can be applied.

4. Dialogue Models and Robot

In our robot, the coordination among modules during the execution of a task is done by the execution of dialogue models. These dialogue models are written using the SitLog programming language for service robots. Dialogue models are composed of a set of *situations* which have associated a tuple of *expectations*, *actions,* and next *situations*. In a typical interaction cycle, the robot is in a given situation in which it has certain expectations; if the arriving interpretation matches some expectation, the robot proceeds to perform the associated actions and moves to another situation. In this framework, performing a task consists in traversing a dialogue model while the modules provide the interpretations and perform the actions defined by the dialogue model. Additionally, in SitLog, it is possible to define the main elements as functions that dynamically construct the structure of the dialogue model. It is also possible to call another dialogue model from a particular situation. Figure 1 depicts these cases:

(a) It shows a typical SitLog arc in which a situation S_i has one expectation α; if this is satisfied, the set of actions β are performed and the arc arrives to the situation S_j.

(b) It exemplifies the case in which the elements are defined by functions and the expectation will be defined by the evaluation of function f, actions by g, and the next situation by h. In particular, these properties make it possible to program dynamic dialogue models which change as the task is executed.

(c) It shows the case for *recursive* situations in which a situation calls an entire dialogue model for execution, and, when finished, it uses the name of the last situation as its own expectation. Algorithm 1 presents the SitLog code for these cases.

SitLog at its core is framework-agnostic; one could implement different frameworks such as subsumption [37], reactive [38], or simple state machine architectures. At our current laboratory, we have developed a full library of behaviours and tasks based on our interaction cycle which is a high level cognitive architecture (IOCA [4, 39]). Associated with

```
[ id ==> s_i,
  arcs ==> [ alpha : betha => s_j]
]
[ id ==> s_i,
  arcs ==> [ f(x) : g(y) => h(z) ]
]
[ id ==> s_i,
  embeded_dm ==> r(m)
  arcs ==> [ alpha : betha => s_j ]
]
```

Algorithm 1: Example of code for SitLog: (a) typical SitLog arc, (b) functional arc, and (c) recursive situation.

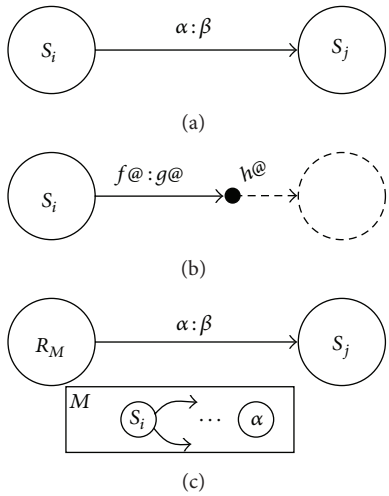


Figure 1: Examples of situations, expectations, and actions. (a) Typical SitLog arc, (b) functional arc, and (c) recursive situation.

the architecture, we have implemented several skills such as vision, language [40], sound [5], navigation, manipulation, and movement. Using this framework, we have programmed the Golem-II+ robot (depicted in Figure 2 [41]). Table 2 summarizes the main capabilities, modules, and hardware that compose the robot. Within this framework and the current version of the hardware, we have implemented several tasks such as following a person, introducing itself, searching for objects, acting as a waiter in a dinner party or a restaurant [42], guarding a museum, and playing a match memory game (for examples of the tasks, visit http://golem.iimas.unam.mx/).

4.1. Multiple-DOA Estimation System. The DOA skill is performed by the multiple-DOA estimation system. This is based on previous work that focuses on a small, lightweight hardware setup that is able to estimate more DOAs than the amount of microphones employed [5, 43].

From the hardware standpoint (see Table 2, sound perception section), the system uses a 2-dimensional 3-microphone array (see Figure 4) going through a 4-channel USB interface. From the software standpoint, the architecture

Figure 2: The Golem-II+ robot.

of this can be seen in Figure 3. The system is divided into three parts.

Audio Acquisition. This is the underlying module that feeds audio data to the next part of the system. It is based on the Jack Audio Connection Toolkit [44], which provides high resolution data (48 kHz, in our case) at real-time speeds with a very high tolerance in the amount of microphones it manages and a relative small resource requirement. This captured audio passes through a VAD which activates the localization.

Initial Single-DOA Estimation. For each time window, each microphone pair is used to estimate a signal delay using a cross-correlation vector (CCV) method (see Figure 3, single-DOA estimation block). Each delay is used to calculate two preliminary DOAs (to consider the back-to-front mirroring issue of 1-dimensional arrays pose). Using a basic search method, the most *coherent* set of 3 DOAs (1 from each microphone pair) is found, using a proposed coherency metric (see Figure 5). If the coherency of that set is above a certain threshold, the DOA of the most perpendicular pair towards the source is proposed as its DOA; this pair is chosen to avoid nonlinearities in the relation between the delay and the resulting DOA. Figure 5 illustrates this stage.

Multiple-DOA Tracking. Because the previous part carries DOA estimation near real-time speeds, it is able to estimate the DOA from one source in small time windows, even in instances where there are two or more simultaneous sound sources (see Figure 3, multi-DOA tracking block). However, the occurrence of 1-sound source time windows in such instances is stochastic. To this effect, when a DOA is proposed from the previous part of the system, a clustering method is

Table 2: Capabilities, modules, and hardware used on the Golem-II+ robot (IH: in-house developed).

Module	Hardware	Software libraries
	Dialogue management	
Dialogue manager	—	SitLog
Knowledge-base	—	Prolog
	Vision	
Object recognition	Flea 3 camera	MOPED
Person recognition	WebCam	OpenCV
Person tracking	Kinect	OpenNI
Gestures recognition	Kinect	OpenNI
	Language	
Speech recognition	Directional Mic	PocketSphinx
Speech synthesiser	Speakers	Festival TTS
Understanding	—	GF grammar, IH parser
	Sound perception	
DOA system	3 omnidirectional mics	—
Volume monitor	Directional mic	Jack, IH library
Volume monitor	M-Audio Fast Track interface	Jack, IH library
	Navigation, manipulation, and neck	
Navigation	Robotic base, laser	Player
Manipulation	IH robotic arm	Dynamixel RoboPlus
Neck movement	IH neck	Dynamixel RoboPlus

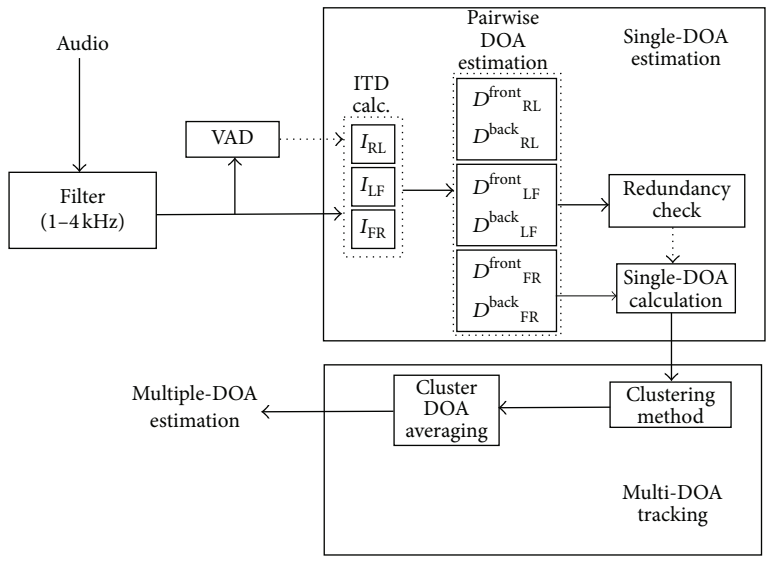


Figure 3: Architecture of the multiple-DOA estimation system.

employed to group several nearby DOAs into one or more estimated sound sources directions.

Figure 6 shows the tracking of three speakers during 30 seconds; each of the speakers is separated by 120 degrees. As it can be seen, this tracking is quite effective, as it could localize each of the found sources. In this example, the system had a 69% precision and 60% recall performance at the frame level. Table 3 presents an evaluation of the whole system. Although it could be thought that this performance would not be enough for HRI interactions, since it is at the frame label (i.e., 100 ms), in an interaction setting when turns are being taken, this performance is more than enough to catch the interaction of a user. An extensive evaluation of this module can be consulted in [43].

5. DOA Behaviour

The goal of the DOA behaviour is to transform DOA measurements into a reference to an object in the context of the task. When the DOA system detects a source, this

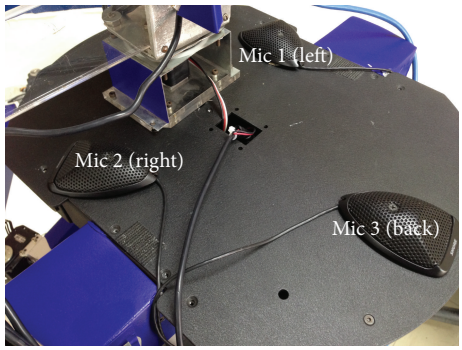


Figure 4: Microphones of the audio localization system in the robot Golem-II+.

Table 3: Multiple-DOA estimation system performance.

Minimum distance of speaker	30 cm
Maximum distance of speaker	5 m
Maximum number of simultaneous speakers	4
Response time	s
Range (minimum angle between speakers)	10
Coverage	360
F1-score 1 speaker	100.00%
F1-score 2 speakers	78.79%
F1-score 3 speakers	72.34%
F1-score 4 speakers	61.02%

sign has a strong relation with the emitter of the sound. The DOA behaviour takes advantage of this strong relation to resolve the object being "pointed at" by imposing a spatial constraint on the possible sources. The spatial constraint consists of the relative position a of the source x given the current robot position (i.e., $\lambda a\ x.rel\_pos(x, a)$). At this point, we only have the information that something is at position a; however, since we consider the DOA an index, we can use our formulation (1) for indices so that we obtain the following reduction:

$$\lambda x.rel\_pos\,(x : a, a)\,. \tag{2}$$

Although we are able to constrain the object, so far, we have not resolved the index to the right referent; to achieve this, we need to use the contextual information. The angle of the DOA is used to identify potential objects given the context. Algorithm 2 is in charge of this resolution stage by analysing which objects on the context satisfy the DOA constraint.

The whole mechanism is encapsulated in a behaviour which is programmed as a simple dialogue model (see Figure 7). This dialogue model can process multiple DOAs at once; for each one, it will try to resolve the object being refered. If successful, it creates an *ok* situation associated with a list of referred objects, otherwise, it reaches an *error* situation. The proposed DOA behaviour supports interaction in two modalities:

(i) Unrestricted call.

(ii) Contextualized call.

```
Require: Pred, DOAs, Context
Ensure: Pred(listObjects, listDOAs)
  listObjects ← []
  listDOAs ← []
  for doa ← DOAs do
    for id, α ← Context do
      if doa α then
        listObjects.push(id)
        listDOAs.push(doa)
      end if
    end for
  end for
```

Algorithm 2: Implementation algorithm for the interpretation of DOAS indices.

The first case corresponds to the scenario in which the position of the user cannot be expected/predicted beforehand; in this case, the description of the context C is empty meaning that only one user is present (Figure 8(a) depicts the robot in such situation); that is, the indices are directly interpreted as the user the robot is having the interaction with. In the second case, the robot has an expectation of the direction from which it will be called and it uses this expectation to identify the caller. For instance, Figure 8(b) depicts the robot having a conversation with two users; with the information of their position, it can discard a third user which is not part of the conversation.

So far, we have assumed that the robot is not moving, but the user should be able to call the robot while it is moving. In order to tackle this situation, we created a composed behaviour which allows the robot to be walking and listening for possible calls at a distance. Figure 9 shows the dialogue model of this behaviour. The call for such behaviour is $walkdoa(D, C)$; D represents the destination and C the description of the spatial context. First, the behaviour starts with the action of *walking* to D and, while doing so, the robot polls for possible DOAs sources in C. If there are none, it continues walking; otherwise, it ends the dialogue with a situation called *interrupted* which contains the information relevant to such interruption. The spatial context gets updated in each call with the current position of the robot. Notice that this behaviour reuses the simple DOA behaviour (situation $R_{doas(C)}$).

For the alternative case, when the user moves, the DOA signal imposes an extra constraint into the index. This constraint is that the DOA at a relative position comes from the same sound source as the previous DOA and this extra information is provided by the tracking stage of our multiple-DOA estimation system. Thanks to this constraint, we can use the same mechanism we have outlined so far to handle this case. In combination with the walking and listening behaviour, they can candle the scenario in which both user and robot move.

Additionally, we also propose to use the DOA behaviour to verify the number of sound sources in a given moment. In particular, we have created a composed behaviour between

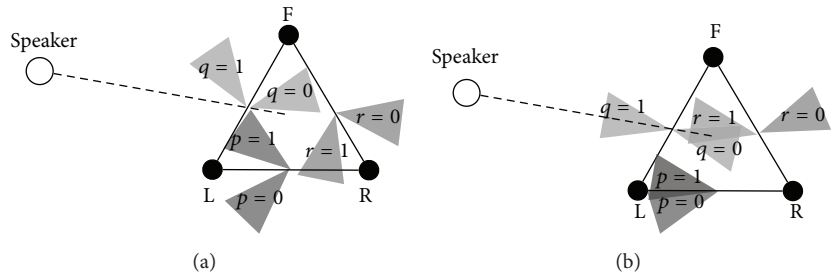


Figure 5: (a) Incoherent set of measures. (b) Coherent set of measures (taken from [5]).

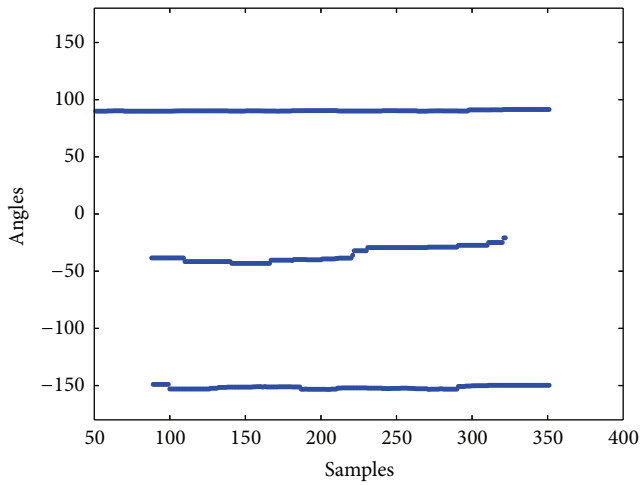


Figure 6: DOAs for three speakers at 90, −30, and −150 degrees.

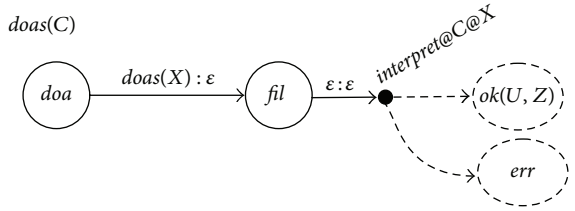


Figure 7: Dialogue model for simple DOA behaviour.

the ask and DOA behaviours which take advantage of this situation. Figure 10 shows the dialogue model for such behaviour. After asking a question P and listening to the answer from the user, the robot checks how many indices were recovered (i.e., the number of elements in Z): if it is only one, it could proceed with the interaction; if they are more than one, the robot infers that more than one user is talking, so it produces an error which has to be handled by the robot. Similarly to the walking behaviour, this behaviour reuses the simple DOA behaviour.

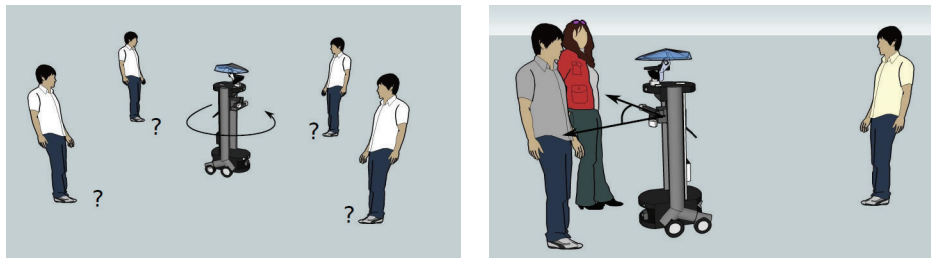
6. Implemented Tasks

The DOA behaviours have been used to program the following tasks with our robot Golem-II+ (videos of these tasks can be watched at https://youtube/Q6prwIjoDnE?list=PL4EiERt__u4faJoxHF1M5EMwhEoNH4NSc).

Following a Person. In this task, the robot follows a person by using the visual tracking system (kinect based). The robot tries to keep a distance of 1 m; while the user is moving, the robot tries to catch it at a safe speed. In case the robot loses him, it asks the user to call it so the robot could infer which direction to look for. Once positioned on the right direction, it uses vision to continue the tracking it uses. When lost, the user can be in any direction and inside a ratio of 5 m and the robot would identify where it is. If the user is not found, it will insist to be called and try again to locate him (for a demo of the full system, see supplementary video *following a person.mp4* in Supplementary Material available online at http://dx.doi.org/10.1155/2016/3081048).

Taking Attendance. In this task, the robot takes attendance in a class by calling the names of the members of the class and for each call it expects an auditory response. When the direction of the response is identified, it faces towards that direction, checks if a person is there by asking to wave, and gets closer to verify the identity of the student. The number of students is limited by the vision system which only can see up to four different persons. They have to be located in a line; at the moment, the task does not consider the case of multiple rows. We take advantage of this setting and we specify a spatial context in which students are only in front of the robot (for a demo of the full system, see supplementary video *taking attendance of a class.mp4*).

Playing Marco-Polo. In this task, the robot plays the role of looking for the users by enunciating *Marco* and expecting one or more *Polo* responses. When the direction of a response is identified, it moves towards that direction if possible. When advancing to this direction, if the laser sensor recognizes that there is someone in front at close distance, it assumes it has won the game; otherwise, it continues calling *Marco* until the user responds. The robot has n tries to catch the user; if not, it gives up and loses the game. The laser hypotheses that there is someone in front if it detects a discontinuity in the reading. In this setting, the players can be located at any direction in a ratio of 5 m (for a demo of the full system, see supplementary video *playing Marco-Polo.mp4*).



(a) (b)

Figure 8: Examples of calls at a distance: unrestricted, one user not at predictable position, and contextualized, multiple users at predictable position.

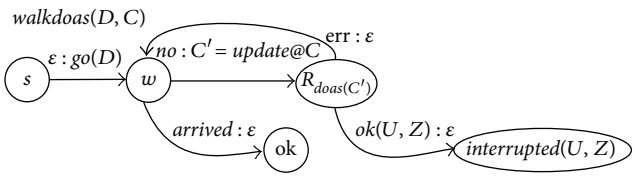


Figure 9: Dialogue model for walk DOA behaviour.

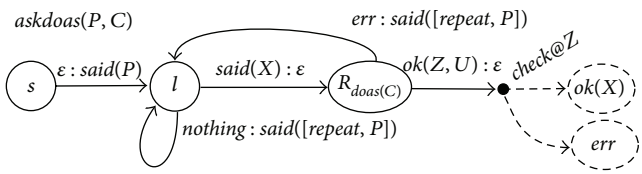


Figure 10: Dialogue model for ask DOA behaviour.

Waiter. In the previous examples, the DOA behaviour relies on the robot's initiative. This task pushes the behaviour into a mixed-initiative strategy. In this case, the robot is a waiter and it waits for calls from the users located in the restaurant tables. Once a table calls at a distance, the robot will walk to the table to take the orders of the clients. However, while walking, it will listen for other calls from different tables; if this happens, it will let the users in the corresponding table know that it will be there as soon as it finishes with the table it is currently walking to. While taking the order, the robot directs the interaction if more than one client talks at the same time. Once it has the order, it proceeds to bring it by taking the drinks/meals from a designated pickup location in the kitchen area or by asking the chef directly for them. This task is limited by the maximum number of sources that could be simultaneously detected (4). In this case, the maximum number of tables to attend would be 4 with a maximum number of 4 clients in each table; this have to be in a ratio of 5 m (for a demo of the whole system, see supplementary video *waiter in a restaurant.mp4*).

We hypothesize that the success of the interaction in each of the tasks is possible because of the use of the DOA information as an index which is fully interpreted in the context of each task. These contexts make the physical relation between the sign (the call) and the object (the user or users) evident. See Table 4 for a summary of the main elements involved in the resolution of the DOA indices in each of the tasks and detailed as follows.

In the case of the *following a person* task, it is until the robot has lost the user and addresses her or him that the response is interpreted as the user and the direction where it could look for its user; this is a case of unconstrained call in which the spatial context is empty. Without context, the robot is able to establish the relative position of the user.

In the case of the *taking attendance* task, the protocol for the interaction is quite standard. The robot calls the name of the student, and she or he has to respond to such call. Such response is interpreted as the presence of the student and its relative position is established. Here, we can constrain the calls to be in front of the robot.

In the case of the *playing Marco-Polo* task, the rules define the context on which the call is interpreted. These rules specify that when the robot says *Marco*, the users have to respond with *Polo*; the index of such response can be interpreted as the relative position of a specific player and it can direct the robot to look in that area. This task also uses an unconstrained call. In fact, the user is not even required to say *Polo*, since the objective of the game is that the robot "catches" one of the users via sound alone; therefore, the DOA of any sound may be considered as an index; this DOA will refer to the user.

In the case of the *waiter* task, the spatial context is defined by the arrangement of the tables in the restaurant. This task uses this spatial context to interpret which table calls for the attention of the robot. The constraint triggered by the DOA information helps identify the calling table, and it also establishes its relative position. The task uses both the simple version and the walking version of the DOA behaviour to resolve the indices. Additionally, it uses the ask version of the DOA behaviour when taking the order. The information provided by the ask version is used by the robot to face the clients while taking their orders and to direct the interaction by asking the users to talk one at the time.

6.1. Application of the DOA Behaviour in Tasks. Figure 11 shows excerpts from the dialogue model of the *following a person*, *taking attendance*, and *playing Marco-Polo* tasks. Given the information provided by the DOA behaviour, the *following a person* and *playing Marco-Polo* tasks cannot

Table 4: Characteristics of the tasks using DOA behaviours.

	Following a person	Taking attendance
Agents	Person being followed	Students
Robot goal	Track a person	Verify students present
Users goals	Being followed	Be in the class
DOA behaviour	Call at a distance	Call at a distance
Situation	User lost	Student name called
Sign	*Here*	*Present*
Expression	$\lambda x.rel\_pos(x : a, a)$	$\lambda x.rel_p os(x : a, a)$
Resolution	$rel_p os(user, a)$	$rel_p os(name, a)$
Interpretation	Target and its position	Presence of student and position
	Playing Marco-Polo	Waiter
Agents	Players	Clients
Robot goal	Catch a player	Take and deliver orders
Users goals	Not being caught	Receive orders
DOA behaviour	Call at a distance	Call at a distance Verification
Situation	Say *Marco*	Robot doing nothing or asking order
Sign	*Polo*	*Waiter*
Expression	$\lambda x.rel\_pos(x : a, a)$	$\lambda x_a.rel_p os(x : a, a)$
Resolution	$rel\_pos(player, a)$	$rel_p os(table, a)$
Interpretation	Player and its position	Table to order and client talking

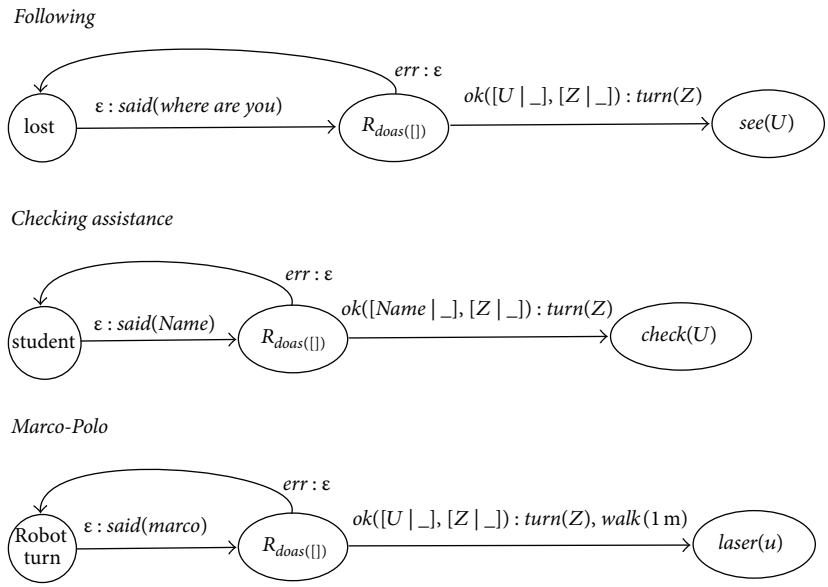


Figure 11: The DOA behaviour used on the *following a person*, *taking attendance*, and *playing Marco-Polo*.

identify the user; it is as if he or she would have only said *I am here*. The response of the robot to this interaction is not verbal but an action; in both cases, it turns towards the direction of the call. In case of an error in the detection of the DOA, the robot will try to recover. In the *following a person* task, the person tracker will trigger an error when no one is found and it will ask again for a call by the user. In the *playing Marco-Polo* task, the robot will repeat *Marco* and wait for the answer, continuing with the game. Because of the nature of the *take attendance* task, the interpretation of the index signals a particular student even in an underspecified context because the robot knows the name of the student being called, and the response is attributed to her or him. It is as if he or she would have responded *me X, I am here*. In this task, the robot tries to prevent an error in the DOA location and it checks if there is an actual person in that direction, but in case it is not, it will keep calling the name of the student.

Figure 12 shows the DOA behaviours for the *waiter* task. This task uses the three behaviours based on the DOA skill. First, it uses the DOA behaviour but with a defined context.

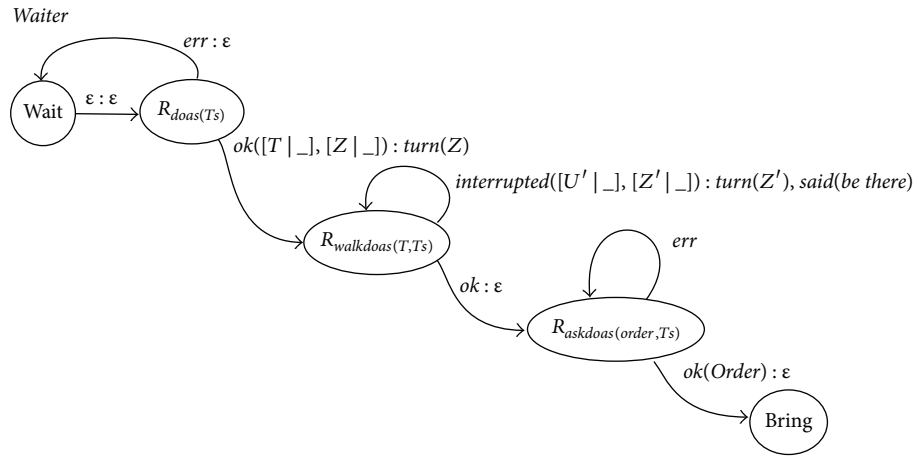


Figure 12: The DOA behaviours used on the *waiter* task.

The interpretation of the index will depend on the spatial context and it will define which table calls the robot. In this case, it is as if the client would have said *here table X needs something*. In the task, an error has a larger effect on the interaction; if the robot mistakenly approaches a table that did not call for it, it will offer to take the order to which the clients there could refuse or if the table is empty the robot will leave. However, the table which really wanted to ask for an order could call the robot while walking, so that when the robot is done with the erroneous table, it will approach the correct one. During such movements, the task uses the walking DOA behaviour. Finally, when arriving to a table to ask an order, the task uses the ask DOA behaviour, and if more than one index is returned during the response, it will let the clients know that only one user is supposed to talk at a time. If an error on the detection of DOAs occurs during this part of the task, it will create a confusing situation on which the robot will face clients which are not there; however, if no client answers, the robot will assume the "ghost" client does not want anything and will carry on with the rest of the order taking and delivery.

6.2. History of Demonstration and Evaluation of the Tasks. The *following a person*, *playing Marco-Polo,* and *waiter* tasks have been repetitively demonstrated in our lab to the general public and in the RoboCup@Home competition [45, 46]. The experience with these demonstrations during the competition has been positive. In 2012, the *following a person* task was demonstrated in the German Open competition, where the demonstration allowed the team to obtain the third place, and in the Mexican Tournament of Robotics national competition, where the robot won the first place. In 2013, the *waiter* task was presented in the first stage of the RoboCup competition and it was awarded the Innovation Award of the @Home league that year. Additionally, that same year, a demonstration of the robot playing Marco-Polo with another robot was attempted but technical problems prevented the execution of the demo during competition, but it was successfully carried out to the general public. In all these cases, we have observed a positive appreciation of the integration of sound localization in the interaction between humans and robots. In particular, the *waiter* task has been extensively evaluated [42]. This evaluation was done with final users (30) which were asked to place an order to the robot. 60% of the users did not have previous experience with a service robot. We found that, when asking for an order, clients repeatedly called at a distance (100%) and fully completed the task (90%). And even though it was not common that two users talked at the same time (40%), when they did, the robot was able to direct the interaction.

In the case of the *taking attendance* task, this was programmed by students during a three-week summer internship. The students had no previous knowledge of our development framework. During the internship, they learned the framework, developed the idea, and implemented it on the robot. Other four teams of students developed other four different tasks. The abstraction of the skill as an element of the SitLog language made it relatively easy to directly implement the task and reach an appropriate performance in a short time span.

7. Discussion

In this work, we proposed the use of direction of arrival of sound sources as indices. In many Human-Robot Interaction tasks, the position of objects is essential for the execution of the task, for instance, losing someone while following him or her. Under this perspective, DOAs are second-class citizens since they do not provide a position but a direction. However, treating DOAs as indices allows the robot to link such directions to objects in the spatial context that leads to the correct interpretation and continuation of the task, guessing that the lost one is a direction. This consideration permits modeling calls at a distance to a robot which is a desirable way to interact with an interactive agent. In order to demonstrate this concept, we implemented a set of behaviours that depend on DOAs to be interpreted as indices and tested them in four tasks: following a person, taking the attendance of a class, playing Marco-Polo, and acting as a waiter in a restaurant.

The spatial context, or a lack of it, defines aspects of the type of interaction on the task. When there is no available spatial context, such as when playing Marco-Polo, any index is considered by the robot as a possible participant. However, when a spatial context is defined, for instance, in attending the tables in a restaurant, it is possible to keep track of more than one user. It is also possible to have a mixed-initiative interaction in which the robot waits for the call, rather than visiting each table and asking if they want anything. The indexicality of the DOAs allows the robot to react to the direction of the user and continue with the interaction, not necessarily in the same modality: in *playing Marco-Polo,* the robot moves in the direction of the index; in *taking the attendance* and *following a person* tasks, the robot confirms the presence of the user by visual methods in the direction of the index; in the *waiter* task, the robot faces the tables which correspond to the indices. These three possible aspects of the interaction when using indexical DOAs (multiuser, mixed-initiative, and multimodal) are an example of the richness of the use of calls at a distance in an interaction.

Additionally, we have showed that the DOA behaviour complements the verbal interaction. In particular, knowing that there is more than one person talking when the robot is listening is a good indication that the interaction may not be good. With the indexical information, the robot can take control of the interaction and direct it by asking each user to speak one at a time, preventing an error in the communication.

We were able to implement these tasks since the multiple-DOA estimation system is light and robust up to four sound sources. However, this system and other robotic systems impose some restrictions on the interaction. In the case of following a person, it can only follow one person at a close distance (1 m); in the case of taking attendance, the students have to be aligned; In our test, there were three students; in the case of playing Marco-Polo and waiting tables, the users, player and tables, have to be in a ratio of 5 m. In all these cases, the maximum number of speakers could have been four. However, under this condition, the performance of the system was the poorest and had an effect on the interaction; our evaluation shows that one user for playing Marco-Polo and two users for table in the waiter case were the limit to achieve good performance. At this point, another major problem is possibility of hijacking the attention of the robot. This is related to the fact that when the spatial context is not defined, the robot automatically interprets the DOA as its target. This is exemplified with the Marco-Polo game. At the moment, our implementation of the task takes advantage of this situation and an index is interpreted as a possible user which becomes the target to approach. However, this situation can provoke the fact that the robot switches targets constantly rather than targeting one which might be a better strategy to win the game. In order to account for this situation, we have identified that the DOA behaviour could be complemented with a speaker verification stage in which the identity of who is responding with *Polo* would be verified. This complement is persistent with our proposal since it becomes an extra constraint in the referred object. Other tasks can also take advantage of this, as such verification can be part of the *following a person*, *taking attendance*, and *waiter* tasks. In the future, we will explore complementing the DOA behaviour with such verification and its consequences on the interaction.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

The authors thank the support of CONACYT through Projects 81965 and 178673, PAPIIT-UNAM through Project IN107513, and ICYTDF through Project PICCO12-024.

References

[1] C. Peirce, N. Houser, and C. Kloesel, *The Essential Peirce: Selected Philosophical Writings*, vol. 1 of *The Essential Peirce*, Indiana University Press, 1992.

[2] C. Peirce, N. Houser, C. Kloesel, and P. E. Project, *The Essential Peirce: Selected Philosophical Writings*, vol. 2 of *The Essential Peirce*, Indiana University Press, 1998.

[3] L. A. Pineda, L. Salinas, I. V. Meza, C. Rascon, and G. Fuentes, "Sit log: A programming language for service robot task," *International Journal of Advanced Robotic Systems*, vol. 10, pp. 1–12, 2013.

[4] L. Pineda, A. Rodrguez, G. Fuentes, C. Rascn, and I. Meza, "Concept and functional structure of a service robot," *International Journal of Advanced Robotic Systems*, vol. 6, pp. 1–15, 2015.

[5] C. Rascon and L. Pineda, "Multiple direction-of-arrival estimation for a mobile robotic platform with small hardware setup," in *IAENG Transactions on Engineering Technologies*, H. K. Kim, S.-I. Ao, M. A. Amouzegar, and B. B. Rieger, Eds., vol. 247 of *Lecture Notes in Electrical Engineering*, pp. 209–223, Springer, Amsterdam, Netherlands, 2014.

[6] S. Argentieri, A. Portello, M. Bernard, P. Danes, and B. Gas, "Binaural systems in robotics," in *The Technology of Binaural Listening*, pp. 225–253, Springer, Berlin, Germany, 2013.

[7] H. G. Okuno and K. Nakadai, "Robot audition: its rise and perspectives," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '15)*, pp. 5610–5614, South Brisbane, Australia, April 2015.

[8] S. Argentieri, P. Danès, and P. Souères, "A survey on sound source localization in robotics: from binaural to array processing methods," *Computer Speech & Language*, vol. 34, no. 1, pp. 87–112, 2015.

[9] R. A. Brooks, C. Breazeal, M. Marjanović, B. Scassellati, and M. M. Williamson, "The cog project: building a humanoid robot," in *Computation for Metaphors, Analogy, and Agents*, vol. 1562 of *Lecture Notes in Computer Science*, pp. 52–87, Springer, Berlin, Germany, 1999.

[10] H. Kitano, H. G. Okuno, K. Nakadai, T. Sabisch, and T. Matsui, "Design and architecture of sig the humanoid: an experimental platform for integrated perception in robocup humanoid challenge," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '00)*, vol. 1, pp. 181–190, IEEE, Takamatsu, Japan, 2000.

[11] F. Michaud, C. Côté, D. Létourneau et al., "Spartacus attending the 2005 AAAI conference," *Autonomous Robots*, vol. 22, no. 4, pp. 369–383, 2007.

[12] 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*, vol. 23, no. 4, pp. 742–752, 2007.

[13] I. Hara, F. Asano, H. Asoh et al., "Robust speech interface based on audio and video information fusion for humanoid HRP-2," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '04)*, vol. 3, pp. 2404–2410, September-October 2004.

[14] J. C. Murray, H. R. Erwin, and S. Wermter, "Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks," *Neural Networks*, vol. 22, no. 2, pp. 173–189, 2009.

[15] H.-D. Kim, J.-S. Choi, and M. Kim, "Human-robot interaction in real environments by audio-visual integration," *International Journal of Control, Automation and Systems*, vol. 5, no. 1, pp. 61–69, 2007.

[16] B.-C. Park, K.-D. Ban, K.-C. Kwak, and H.-S. Yoon, "Sound source localization based on audio-visual information for intelligent service robots," in *Proceedings of the 8th International Symposium on Advanced Intelligent Systems (ISIS '07)*, pp. 364–367, Sokcho, South Korea, 2007.

[17] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '09)*, pp. 2033–2038, St. Louis, Miss, USA, October 2009.

[18] D. Bohus and E. 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 (ICMI-MLMI '10)*, pp. 5:1–5:8, Beijing, China, November 2010.

[19] K. Nakadai, H. G. Okuno, and H. Kitano, "Real-time sound source localization and separation for robot audition," in *Proceedings of the IEEE International Conference on Spoken Language Processing (ICSLP '02)*, pp. 193–196, Denver, Colo, USA, September 2002.

[20] V. M. Trifa, A. Koene, J. Morén, and G. Cheng, "Real-time acoustic source localization in noisy environments for human-robot multimodal interaction," in *Proceedings of the 16th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN '07)*, pp. 393–398, August 2007.

[21] J. G. Trafton, M. D. Bugajska, B. R. Fransen, and R. M. Ratwani, "Integrating vision and audition within a cognitive architecture to track conversations," in *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI '08)*, pp. 201–208, ACM, March 2008.

[22] K. Teachasrisaksakul, N. Iemcha-Od, S. Thiemjarus, and C. Polprasert, "Speaker tracking module for indoor robot navigation," in *Proceedings of the 9th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON '12)*, 4, 1 pages, May 2012.

[23] X. Li, M. Shen, W. Wang, and H. Liu, "Real-time Sound source localization for a mobile robot based on the guided spectral-temporal position method," *International Journal of Advanced Robotic Systems*, vol. 9, article 78, 2012.

[24] B. Fransen, V. Morariu, E. Martinson et al., "Using vision, acoustics, and natural language for disambiguation," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction ((HRI '07)*, pp. 73–80, Arlington, Va, USA, March 2007.

[25] K. Nakadai, S. Yamamoto, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "A robot referee for rock-paper-scissors sound games," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '08)*, pp. 3469–3474, IEEE, Pasadena, Calif, USA, May 2008.

[26] I. Nishimuta, K. Itoyama, K. Yoshii, and H. G. Okuno, "Toward a quizmaster robot for speech-based multiparty interaction," *Advanced Robotics*, vol. 29, no. 18, pp. 1205–1219, 2015.

[27] H. M. Do, W. Sheng, and M. Liu, "An open platform of auditory perception for home service robots," in *PRoceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '15)*, pp. 6161–6166, IEEE, Hamburg, Germany, September-October 2015.

[28] A. G. Brooks and C. Breazeal, "Working with robots and objects: revisiting deictic reference for achieving spatial common ground," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction (HRI '06)*, pp. 297–304, Salt Lake City, UT, USA, March 2006.

[29] O. Sugiyama, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "A model of natural deictic interaction," *Human-Robot Interaction in Social Robotics*, vol. 104, 2012.

[30] H. G. Okuno, K. Nakadai, K.-I. Hidai, H. Mizoguchi, and H. Kitano, "Human–robot non-verbal interaction empowered by real-time auditory and visual multiple-talker tracking," *Advanced Robotics*, vol. 17, no. 2, pp. 115–130, 2003.

[31] A. Atkin, "Peirce's theory of signs," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., 2013.

[32] D. Kaplan, "Dthat," in *Syntax and Semantics*, P. Cole, Ed., vol. 9, pp. 221–243, Academic Press, New York, NY, USA, 1978.

[33] L. Pineda and G. Garza, "A model for multimodal reference resolution," *Computational Linguistics*, vol. 26, no. 2, pp. 139–193, 2000.

[34] R. Mitkov, *Anaphora Resolution*, vol. 134, Longman, London, UK, 2002.

[35] J. R. Tetreault, "A corpus-based evaluation of centering and pronoun resolution," *Computational Linguistics*, vol. 27, no. 4, pp. 507–520, 2001.

[36] V. Ng and C. Cardie, "Improving machine learning approaches to coreference resolution," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 104–111, Association for Computational Linguistics, 2002.

[37] R. A. Brooks, "How to build complete creatures rather than isolated cognitive simulators," in *Architectures for Intelligence*, K. VanLehn, Ed., pp. 225–239, 1991.

[38] R. P. Bonasso, R. J. Firby, E. Gat, D. Kortenkamp, D. P. Miller, and M. G. Slack, "Experiences with an architecture for intelligent, reactive agents," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 9, no. 2-3, pp. 237–256, 1997.

[39] L. A. Pineda, I. Meza, H. H. Avilés et al., "IOCA: an interaction-oriented cognitive architecture," *Research in Computer Science*, vol. 54, pp. 273–284, 2011.

[40] I. Meza, C. Rascon, and L. A. Pineda, "Practical speech recognition for contextualized service robots," in *Advances in Soft Computing and Its Applications: 12th Mexican International Conference on Artificial Intelligence, MICAI 2013, Mexico City, Mexico, November 24–30, 2013, Proceedings, Part II*, vol. 8266 of *Lecture Notes in Computer Science*, pp. 423–434, Springer, Berlin, Germany, 2013.

[41] L. Pineda and G. Golme, "Grupo Golem: robocup@home," in *Proceedings of the RoboCup*, 8 pages, RoboCup Federation, Eindhoven, The Netherlands, June 2013.

[42] C. Rascon, I. Meza, G. Fuentes, L. Salinas, and L. A. Pineda, "Integration of the multi-doa estimation functionality to human-robot interaction," *International Journal of Advanced Robotic Systems*, vol. 12, no. 8, 2015.

[43] C. Rascon, G. Fuentes, and I. Meza, "Lightweight multi-DOA tracking of mobile speech sources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–16, 2015.

[44] P. Davis, JACK Connecting a World of Audio, http://jackaudio.org.

[45] T. van der Zant and T. Wisspeintner, "RoboCup X: a proposal for a new league where RoboCup goes real world," in *RoboCup 2005: Robot Soccer World Cup IX*, A. Bredenfeld, A. Jacoff, I. Noda, and Y. Takahashi, Eds., vol. 4020 of *Lecture Notes in Computer Science*, pp. 166–172, 2006.

[46] T. Wisspeintner, T. van der Zant, L. Iocchi, and S. Schiffer, "RoboCup@Home: scientific competition and benchmarking for domestic service robots," *Interaction Studies*, vol. 10, no. 3, pp. 392–426, 2009.