This document is published in:
"Cybernetics and Systems: An International Journal, 2013, vol. 44 (8), 681-703. ISSN: 1087-6553. DOI: http://dx.doi.org/10.1080/01969722.2013.832096".



This work has been supported by the CAM Project S2009/DPI-1559/ROBOCITY2030 II, developed by the research team RoboticsLab at the University Carlos III of Madrid.

# Multimodal Fusion as Communicative Acts during Human–Robot Interaction

FERNANDO ALONSO-MARTÍN, JAVIER F. GOROSTIZA, MARÍA MALFAZ, and MIGUEL A. SALICHS

*RoboticsLab, Carlos III University of Madrid, Leganés, Madrid, Spain*

*Research on dialog systems is a very active area in social robotics. During the last two decades, these systems have evolved from those based only on speech recognition and synthesis to the current and modern systems, which include new components and multimodality. By* multimodal dialogue *we mean the interchange of information among several interlocutors, not just using their voice as the mean of transmission but also all the available channels such as gestures, facial expressions, touch, sounds, etc. These channels add information to the message to be transmitted in every dialogue turn. The dialogue manager (IDiM) is one of the components of the robotic dialog system (RDS) and is in charge of managing the dialogue flow during the conversational turns. In order to do that, it is necessary to coherently treat the inputs and outputs of information that flow by different communication channels: audio, vision, radio frequency, touch, etc. In our approach, this multichannel input of information is temporarily fused into communicative acts (CAs). Each CA groups the information that flows through the different input channels into the same pack, transmitting a unique message or global idea. Therefore, this temporary fusion of information allows the IDiM to abstract from the channels used during the interaction, focusing only on the message, not on the way it is transmitted. This article presents the whole RDS and the description of how the multimodal fusion of information is made as CAs. Finally, several scenarios where the multimodal dialogue is used are presented.*

*KEYWORDS communicative acts, HRI, human–robot interaction, multimodal dialog, multimodal fusion, robot dialog system*

Address correspondence to Fernando Alonso Martín, RoboticsLab, Carlos III University of Madrid, 28911 Leganés, Madrid, Spain. E mail: fernando.alonso@uc3m.es

# INTRODUCTION

The belief that humans will be able to interact with machines through natural conversation has been one of the favorite issues in science fiction for a long time. Due to the recent advances in the natural interaction field, using voice, gestures, etc., this beginning to become a reality. Dialogue, understood as the information exchange mechanism used by humans, is the most natural way to execute a number of daily actions, such as obtaining information, hiring a service, etc.

Dialogue management is not a new research area and there are many approaches to the natural dialogue problem (not only in robotics). Many researchers have focused on this problem, such as Black et al. (2010), Bohus and Rudnicky (2009), Gorostiza and Salichs (2010), Larsson and Tram (2000), Reithinger and Alexandersson (2003), Wahlster (2003), Walker et al. (2001), and Yin (2010). Nevertheless, the most developed dialogue systems have been implemented in very restrictive communicative contexts, such as the telephone or access services through PC. The implementation of dialogue systems in robotics is not at the same level, mainly due to the fact that robots have a physical body and they move in real environments. Therefore, the dialogue systems must be coordinated with the perception and actuator systems in a real environment, which is a much more complex task than those related to customer support services or smartphone control.

In fact, the multimodality of the dialogue systems is much more interesting for social robotics due to the sensitive and expressive features of these robots. In this case, the number of sensors and actuators is much higher than in any other system (Waibel and Suhm 1997; Cheyer et al. 1998; Niklfeld and Finan 2001; Wahlster 2003; Gorostiza et al. 2006a; Seneff et al. 1996).

In general, a multimodal system supports communication with the user through different modalities such as voice, gesture, nonverbal sounds, typing, etc. (Nigay and Coutaz 1993). Literally, multi refers to more than one and the term modal may cover the notion of modality as well as that of mode. Modality refers to the type of communication channel used to convey or acquire information. It also covers the way an idea is expressed or perceived or the manner in which an action is performed. Our definition of multimodality conveys two salient features that are relevant to the software design of multimodal systems: the fusion of different types of data from/to different input devices and the temporal constraints imposed on information processing from input devices.

In order to understand the concepts of dialog and multimodality it is necessary to understand the communicative acts (CAs) theory. This theory is derived from the study of the verbal communication among humans (Bruner 1975; Searle 1969, 1975; Bach and Harnish 1969), where the CAs are clearly defined as the basic units of the dialogue. Based on those studies,

engineers have found that this theory is very useful for describing and formalizing the communication between humans and machines. The CA represents the message to be transmitted by each interlocutor in each turn during a conversation.

A conversation between a user and a robot is carried out using turns, which involve verbal and nonverbal communication. That is, the user tries to transmit a message by using his or her voice, by adopting a spatial position in relation to the robot, by gestures, by facial expressions, etc. During each of those turns, some information is given, which is important for the dialogue, and it is necessary to group it (into a CA) in order to have a coherent meaning. Then, this information is extracted and managed by the dialogue manager (IDiM).

In Figure 1 an example of the exchange of information during a dialogue between a user and a robot is shown. In each turn, the user and the robot try to transmit a message using different channels. The first temporal line in the figure represents the CAs of the user, the second the CAs of the robot, and the third the dialogue system process. First, the user communicates with the robot through the first CA and the dialogue system takes a certain amount of time, normally less than one second, to process the information and to prepare the answer. Then, the robot answers with its first CA, and the user receives the information transmitted by the robot; but before the CA finishes, the user interrupts the robot with a new CA. Therefore, the robot suspends its CA and keeps quiet until the user finishes his or her exposition. In this sense, the dialogue system is full-duplex, because both interlocutors can interrupt their communicative processes.

The objective of this work is to present how the multimodal fusion is made in the robotic dialogue system (RDS). The RDS is introduced in order to understand and contextualize the fusion of the information received through the different input channels in CAs for the IDiM. Such multimodal fusion allows the system to abstract from the channels used in every dialogue

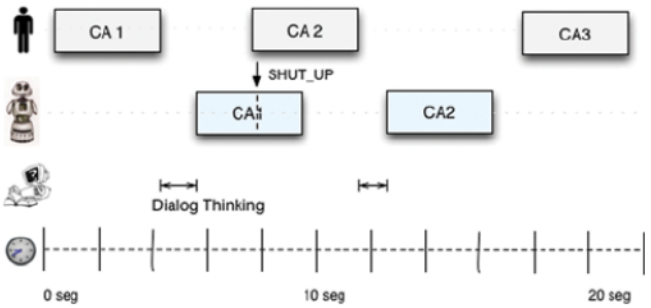


**FIGURE 1** Dialogue with communicative acts (color figure available online).

turn. In order to understand these fusion mechanisms and the RDS, some examples of dialogues are presented at the end of the article.

This article is structured as follows: The following section presents a brief review of some related works, in which multimodal dialogue systems have been implemented on social robots. Next, the experimental platform, the social robot Maggie, is described according to its software and hardware. In the next section, the RDS is introduced and its principal features are also described. Later, we describe the process of the multimodal fusion of information into CAs for the RDS. Next, some scenarios where the multimodal fusion process can be analyzed are presented. Finally, conclusions and future works are outlined.

## RELATED WORK

There is not much literature on the implementation of multimodal dialogue systems in human–robot interaction (HRI). In this section, we briefly describe the most relevant dialogue systems applied to HRI to fulfill the multimodal fusion of the information received by the input channels.

In Fry et al. (1998), a dialogue system was implemented on the Jijo-2 robot. In this work, the communication channel used was just the voice. In more advanced systems, multimodality was also introduced, as shown in Cassimatis et al. (2004), Perzanowski et al. (2001), and Lemon et al. (2002), among others. In these works the verbal information of the user was fused with tactile gestures received from a tactile screen, which shows some type of information (a map or a menu). These systems are able to solve utterances such as "go there" (while pointing at a specific location of the screen). In fact, in Lemon et al. (2002) the visual interface was also used by the dialogue system to show videos while the system generates language. For example, it asks "Is this the red car we are looking for?" while it shows a video of what the robot's camera is detecting.

In Hüwel et al. (2006) and (Toptsis et al. 2004) the BIRON robot is presented. Its dialogue system fuses together the visual information of the user pointing (a certain direction or at an object) with verbal deixis. Therefore, it is able to resolve utterances such as "this is a plant" (while pointing at a plant).

Moreover, multimodal fusion has been used in several works for natural programming systems. For instance, in Iba et al. (2002), a vacuum robot is programmed using both gestures and speech, so the robot is able to fuse multimodal information and to infer the action the user wants to activate. Similar works can be found in Dominey et al. (2007), where the HRP-2 robot fuses visual and verbal information for object manipulation.

In another interesting work, Stiefelhagen (2004) presented a multimodal dialogue in HRI that uses speech, head pose, and gestures. The components are a speech recognizer, 3D face and hand tracking, pointing gesture

recognition, head pose recognition, a dialogue component, speech synthesis, a mobile platform, and a stereo camera system. Three years later, Stiefelhagen and Ekenel (2007) presented their system for spontaneous speech recognition, multimodal dialogue processing, and visual perception of a user, which included localization, tracking, and identification of the user; recognition of pointing gestures; as well as the recognition of a person's head orientation. Each of the components is described in the paper and experimental results are presented. They also presented several experiments on multimodal HRI, such as interaction using speech and gestures, as well as interactive learning of dialogue strategies. The work and the components presented constitute the core building blocks for audiovisual perception of humans and multimodal HRI. These are used for the humanoid robot developed within the German research project Sonderforschungsbereich on humanoid cooperative robots.

Recently, many efforts have been made toward the development of multimodality at the inputs and outputs of the system: multimodal symmetry. A dialogue system with multimodal symmetry implies that the multimodality is present at the inputs as well as at the outputs of the system. This does not imply that there are the same number of input channels as output channels. The management of the multimodal inputs is known as multimodal fusion, and the management of the multimodal outputs is known as multimodal fission.

The SMARTKOM project (Reithinger and Alexandersson 2003); (Wahlster 2003) introduced a very powerful dialogue system in which one important feature is multimodal symmetry. Its novelty is not only the multimodality at the inputs and outputs but also the management mechanisms, mutual disambiguation, multimodal inputs synchronization, and resolution and generation of ellipsis and multimodal anaphora. Nevertheless, its main application is not focused on robotics but in airports, train stations, hotels, and public places, in general.

Our dialog system, called the robotic dialogue system, is conceived and implemented as a system with multimodal symmetry, similarly the one presented in the SMARTKOM project but applied to social robotics. In this field, we introduce a new concept in our system—*multisound*—which will be explained in the following sections.

## OUR SOCIAL ROBOT

Maggie (see Figure 2), is a social robot used as a research platform for HRI studies (Salichs et al. 2006). Our research focuses on finding new ways of improving robots in order to provide the user with new ways of working, learning, and having fun with them.

In this work, the multimodal dialogue system of Maggie is presented. Maggie communicates with users through different input and output channels. Next, those channels will be briefly described.

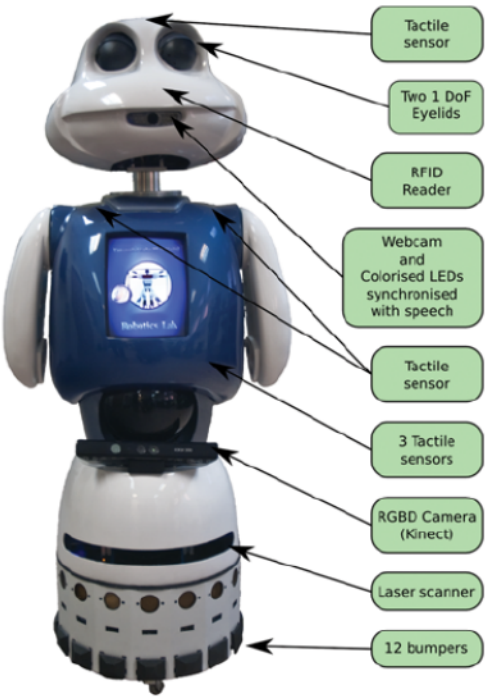


**FIGURE 2** Maggie, our social robot (color figure available online).

## Input Channels

The social robot Maggie has several input channels that are essential for building a multimodal system. Each of them allows the transmission of certain kinds of information, which can be complemented by that received from the rest of the channels:

- *Audio*. Audio is received through a microphone system onboard Maggie. Eight microphones, built into the robot, form a circumference of 40 cm radius at 21 cm from the floor. They are connected by USB ports to the internal computer of the robot and are mainly used for sound source localization tasks. For speech analysis purposes, the robot have two other microphones in its head. These microphones are provided with noise cancellation mechanisms by hardware and echo cancellation by software. Noise cancellation removes uniform noises, such as fan noise; echo cancellation avoids the coupling of the sounds generated by the robot with the inputs of its own microphones. In situations where the environment is very noisy, it is also possible to interact with the robot using wireless or bluetooth earphones. In this case, the interaction will not be very natural. The sound captured by the robot is mainly used for the following tasks: speech recognition based on grammar, speech recognition based on plain text, emotion

**FIGURE 3** RFID tags in Maggie (color figure available online).

detection, user localization, arousal level calculation, sound generation, and musical accompaniment.

- *Vision.* The robot has three basic mechanisms for perceiving its environment: a web camera inside its mouth,[1] a 3D camera Kinect,[2] and a laser telemeter above its base.[3] These sensors are in charge of several tasks: navigation, user detection and identification, gesture and pose detection, object detection, and written text reading based on optical character recognition (OCR).
- *Radio-frequency identification.* Maggie is also able to read radio-frequency identification (RFID) tags. For that purpose, it has several readers onboard: two short-range readers (about 20 cm), one of them in its head and the other in its base, and two long-range readers (about 100 cm) at its sides. The interaction through RFID tags is mainly focused on the identification of objects, such as medicines. In fact, there is another application of these RFID tags that is oriented to the dialogue control. During this dialogue, each tag represents, through drawings, different skills the robot has (to follow someone, to read the news, etc.); see Figure 3. This kind of interaction is more suitable for small children, elderly people, or noisy environments where the interaction through voice can be difficult.
- *Touch.* The robot is endowed with several capacitive sensors that are able to detect whether the user is touching it. The capacitive sensor is not able to detect different touch pressures; it can only determine whether the robot has been touched or not. These sensors are also used to simulate tickling.

[1] http://www.logitech.com/es es/webcam communications/webcams

[2] http://www.xbox.com/es es/kinect

[3] http://www.sick.com/group/EN/home/products/product portfolio/optoelectronic protective devices/Pages/safety laser scanners.aspx

- *Smartphones and tablets*. It is also possible to enter information through the tablet situated in Maggie's chest or through smartphones. In both cases, several sets of options (depending on the kind of dialogue and the objective of the interaction) are presented, and the user can activate or deactivate them using his or her fingers.

### Output Channels

Since we are presenting a multimodal symmetry dialogue system, the multimodality also affects the output channels of the robot:

- *Audio*. The robot has four loudspeakers at the bottom of its head, which allow communication between the robot and the user through voice and sounds. These loudspeakers are used for emotional voice generation, non-verbal sounds, music generation, and music playing.
- *Engagement gestures*. Using its arms, head, eyelids, and mobile base, the robot is able to make gestures that complement the dialogue. Among the repertory of gestures, we can find the following: to say "no" and "yes" with the head, to dance, etc. These gestures are synchronized with the voice and sounds through the IDiM.
- *Infrared sensors*. Maggie is able to control electrical appliances using infrared communication. For that purpose, the robot is endowed with a programmable infrared remote control that emits an adequate signal to switch some devices on/off, such as the TV or the air conditioning.

### AD, the Control Software Architecture

Maggie has a control software architecture developed by the RoboticsLab[4] research group: the automatic-deliberative (AD) architecture. This architecture has two levels: the automatic level and the deliberative level. The automatic level is in charge of the low-level control actions. In addition, the control primitives, which provide the communication and control of the hardware (sensors and motors), are set in this same level. On the other hand, the reasoning abilities and the decision-making capabilities are carried out by the deliberative level.

The main component of the AD architecture is the "skill." A *skill* is an entity with reasoning and information processing capabilities and also capable of carrying out actions, conceptually similar to a Node in ROS.[5] Moreover, one skill is able to communicate with other skills at the same time. A more detailed description of the AD architecture can be found in Salichs et al. (2006), Barber and Salichs (2002), and Gorostiza et al. (2006b).

[4]http://roboticslab.uc3m.es/roboticslab/
[5]http://www.ros.org/wiki/

## THE ROBOTIC DIALOGUE SYSTEM (RDS)

### Components of the RDS

The dialogue system proposed in this work needs a time-coherent input of information that is obtained by fusing into a CA the inputs received by each of the input channels. The dialogue system is composed of many components, as can be observed in Figure 4, and each of them carries out specific tasks in an uncoupled and distributed way. These components are the following:

1. *Dialogue manager*: This component (IDiM) is in charge of managing the dialogue turns and, given certain inputs, generating outputs. This manager is based on the slot-filling paradigm, specifically based on the Voice XML 2.1[6] standard but extended with multimodal functions. Other authors (Lucas 2000; Niklfeld and Finan 2001; Bennett et al. 2002, Eberman et al. 2002; Kibria and Hellström 2007), also used this same interaction system, although none of these works has been implemented on social robots but on telephone or web environments.
   As shown in Figures 4 and 5, the IDiM needs the information stored in two XML files. One of them is based on the Voice XML standard, and it is where the dialogue specifications are written (by the programmer) following the slot-filling paradigm. The second one stores the fused multimodal information (the CA). In order to represent the multimodal information in a coherent way, the W3C organization[7] has created a standard that specifies, through this XML file, how this information must be structured. This standard is called NLSML Natural Language Semantics Markup Language[8] (NLSML) and establishes the rules and the structure that this XML file must have. The fused information received from the different input channels is stored here so the IDiM can manage it.
   Therefore, the IDiM tries to fill one slot of information before passing to the next one, using the XML file that stores the information of the CA. A simple example of this paradigm would be the flight booking process, in which the slots of information to be filled would be the time departure, the city of departure, and the airline. All of these slots can be filled by several questions/answers between the user and the system or through a unique sentence, such as: "I would like to leave Madrid at 7 in the morning and land in Paris at 9 in the morning with Ryanair."

[6] http://www.w3.org/TR/voicexml21/
[7] http://www.w3.org/
[8] http://www.w3.org/TR/nl spec/

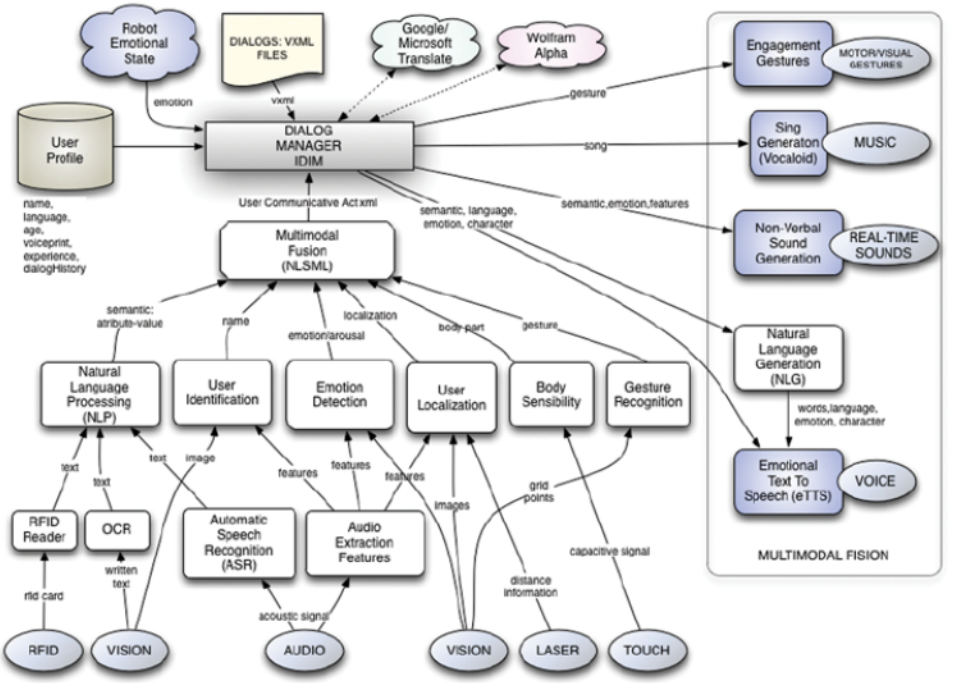


**FIGURE 4** Robotic dialogue system (RDS) (color figure available online).

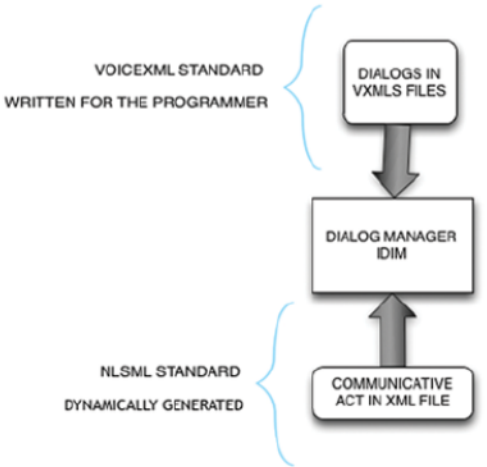


**FIGURE 5** XML files for the dialogue manager (color figure available online).

2. *Input modules*: These are the components that work with the sensory inputs in order to process the perceived information. These components are the following:
- *Automatic speech recognizer (ASR)*: It is in charge of translating voice into text. For this purpose, a modular system is implemented, which allows the connection of several voice recognizers to work in parallel over the same audio samples. Currently, we are working on the voice recognizer based on grammar by Loquendo and on the recognizer based on the Google ASR web service.
- *Natural language processing*: Works together with the ASR for extracting the semantic meaning of the recognized text.
- *Hand-written text to machine-written text converter*: Converts hand-written text into valid text for the dialogue system. In order to do so, OCR techniques are used. See Alonso-Martin et al. (2011) for more details.
- *RFID tag reading*: Reads the information written on active or passive RFID tags; see Corrales and Salich (2009).
- *User identification*: This component is in charge of the identification of the user who is interacting with the robot by his or her voice tone. For this purpose, during the user registration phase, the system saves the user's voice tone (voiceprints).
- *User emotion identification*: user's emotions are extracted from the perceived voice tone of the user. In order to do this, it is necessary to construct a classifier for determining which emotion of the user corresponds with the extracted features of his or her voice, the voiceprints, and his or her previous experience with the dialogue.
- *User localization*: This component localizes the sound source using a complex auditive system formed by eight microphones. The sound source localization system is complemented by laser information for determining the position of the user with greater precision.
- *Touch system*: This system is able to detect whether a robot limb has been touched. Currently, this component is not able to determine the pressure applied by the user.
- *User pose detection*: This component determines the user pose and is able to determine whether the person is sitting, standing, or pointing to the left/right/front, among others. This system uses a deep stereoscopic camera and automatic learning mechanisms.
- *Others*: There are other components that facilitate the introduction of information through tablets, smartphones, or joypads.

3. *Output modules*: These are the modules in charge of expressing or transmitting the message given by the IDiM to the user and work with the outputs of the system. These components are as follows:
- *Emotional text to speech (eTTS)*: This model allows the execution of complex tasks, such as managing the sentences queue, translating texts

into more than 40 languages, and adopting different voice tones according to the system used: Loquendo, Festival, Microsoft TTS, and Google TTS. The "nonverbal sound generation" module makes it possible to synthesize nonverbal sounds in real time, which allows the robot to express messages in its own "robotic" language. These kinds of sounds are generated through the musical programming language Chuck. The "Sing Generation" module allows the robot to sing using the Vocaloid musical software.

- *Natural language generation (NLG)*: The NLG system converts a computer-based representation into a natural language representation. Our NLG is a basic system that converts some semantic values into concrete sentences. For example, "Greet" could be converted into "Hello, I am glad to talk to you" or "Hi, mate."
- *Engagement gestures*: This component allows the robot to carry out typical gestures executed during a normal conversation, using its arms, eyelids, head, neck, and base.

4. *Other modules*: There are other components present in the dialogue system that communicate with web services and give the information needed by the dialogue manager. These services used are automatic translators (Google[9] and Microsoft Translate[10]), a semantic searcher (Wolfram Alpha[11]), and a musical player, Goear.[12]

## Main Features of the Robotic Dialogue System

Although the description of the RDS is not the objective of this article, it is convenient to introduce its main characteristics in order to understand how it works:

- *Interpreted*: The system is interpreted, that is, the dialogue specification of each situation (described in XML files by the programmer) is uncoupled from its interpretation and execution by the IDiM. Therefore, there are two parts: the dialogue specified in XML files, which establishes certain slots of information to fill, and the software part, which is the IDiM that interprets and executes the dialogues.
- *Adaptable*: The dialogue can be adapted to every user based on static and dynamic features. The static ones are stored in the user profiles that are learned during the HRI through natural dialogue: language, name, age, and voiceprints. The dynamic factors of the user, such as the experience with the system, the emotion detected, the spatial situation of the robot

[9] http://translate.google.es/
[10] http://www.microsofttranslator.com/
[11] http://www.wolframalpha.com/
[12] http://www.goear.com/[13]

(proxemics), etc., are also useful to personalize the interaction. This adaptability is also reflected in the multilanguage: it is possible to establish the dialogue in several languages. In order to do this, the system relies on several on-line translators.

- *Multimodal symmetry*: The interaction can be carried out using several input and output channels. In this sense, the multimodality is taken into account at the inputs (multimodal fusion) as well as at the outputs (multimodal fission) of the system.
- *Multisound*: By multisound we mean that the system is able to manage sound information other than voice. The system is capable of analyzing the input sounds for different skills, such as user localization, emotional classification of the user voice, environmental arousal level detection, voice recognition, and user identification. Moreover, the system is able to synthesize sounds for emotions-by-voice generation, robotic nonverbal sound generation, singing (musical expression), and playing music on-line.

## MULTIMODAL FUSION PROCESS DESCRIPTION

This work is focused on the fusion of the input information that is given to the IDiM. This process is carried out by the multimodal fusion module. There are two levels in relation to the multimodal fusion process: the low and the high level, as shown in Figure 6. The low level is in charge of the fusion of the information received by the perception or input modules. For example, in order to execute the user localization task, we need to fuse the information of the audio and vision inputs. At this level, the information is fused in order to obtain a more relevant information for the dialogue. At the other level, the multimodal fusion is made at a higher level of abstraction. All of the relevant information for the dialogue, provided by other modules, must be temporarily fused into a CA and stored in an XML file and given to the IDiM. In this sense, the high-level multimodal fusion is related more to temporal aspects and CAs identification than to the different input channels (Shimokawa and Sawaragi 2001; Falb et al. 2007).

Therefore, the main idea is that it is possible to communicate with the dialogue system through different input channels: voice, touch, gestures, RFID tags, sounds, etc. The information transmitted simultaneously by each channel is grouped by the multimodal fusion module, forming a CA that embraces all of them. This CA allows the IDiM to abstract from the input channels used. The information given by the CA is used by the IDiM to fill the slots of information in every dialogue.

In order to fill the slots of information, the multimodal fusion module must know how to fuse the information into CAs. In order to do that, it is necessary to establish the beginning and the end of each CA. During each

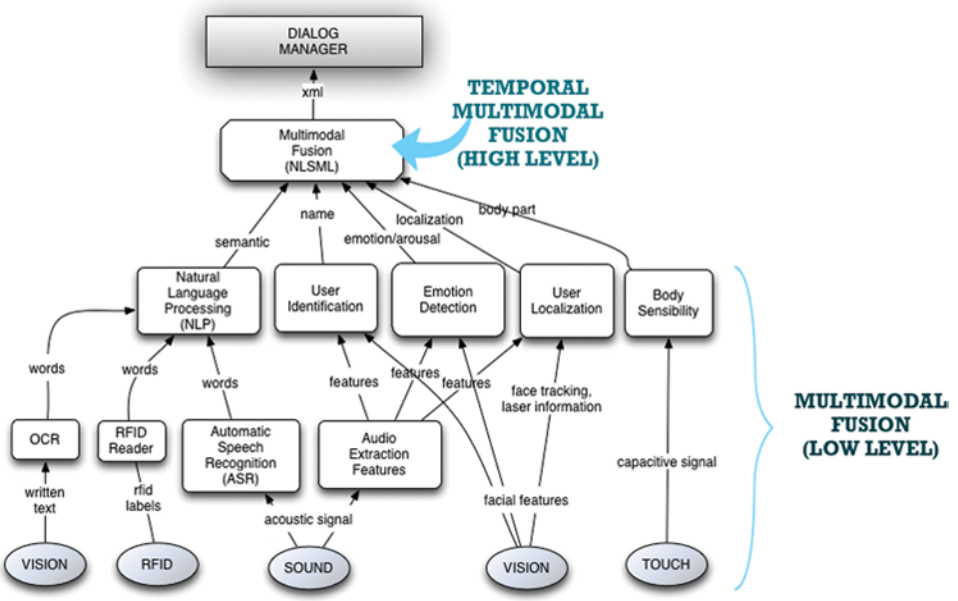


**FIGURE 6** Multimodal fusion levels (color figure available online).

dialogue, each input module sends an event corresponding to the moment when its information is given to the multimodal fusion module. Analyzing the moments when those events arrive, it has been determined that each CA is separated from the others by the notification of the end of the voice recognition. That is, the event given by the ASR when the voice recognition has finished is used to tag the frontier between different CAs. In the case of having several recognition engines running in parallel (Loquendo and Google ASR), the system must wait for both notifications to conclude the current CA.

In Figure 7 a temporal sequence of the multimodal fusion process to create a CA is shown. Each of the module outputs is given to the multimodal fusion module as soon as they are available. This implies that the reception of values is an asynchronous process, the multimodal fusion module being the one that groups them into the same CA. As previously stated, this CA is stored in an XML text file, which follows the NLSML standard, and given to the IDiM. In this figure, two CAs are represented. In the first one, the user touches the robot's right arm and orders it to raise it. In the second one, the user touches the robot's head and orders it to turn left. These are two very simple examples to explain how the multimodal fusion is made.

In the case where the same module launches several events during the same CA, only the last one received is fused into this CA. As shown in Figure 7, if the user touches the head and later the arm during the same dialogue turn, only the arm-touching event is fused into the corresponding CA.

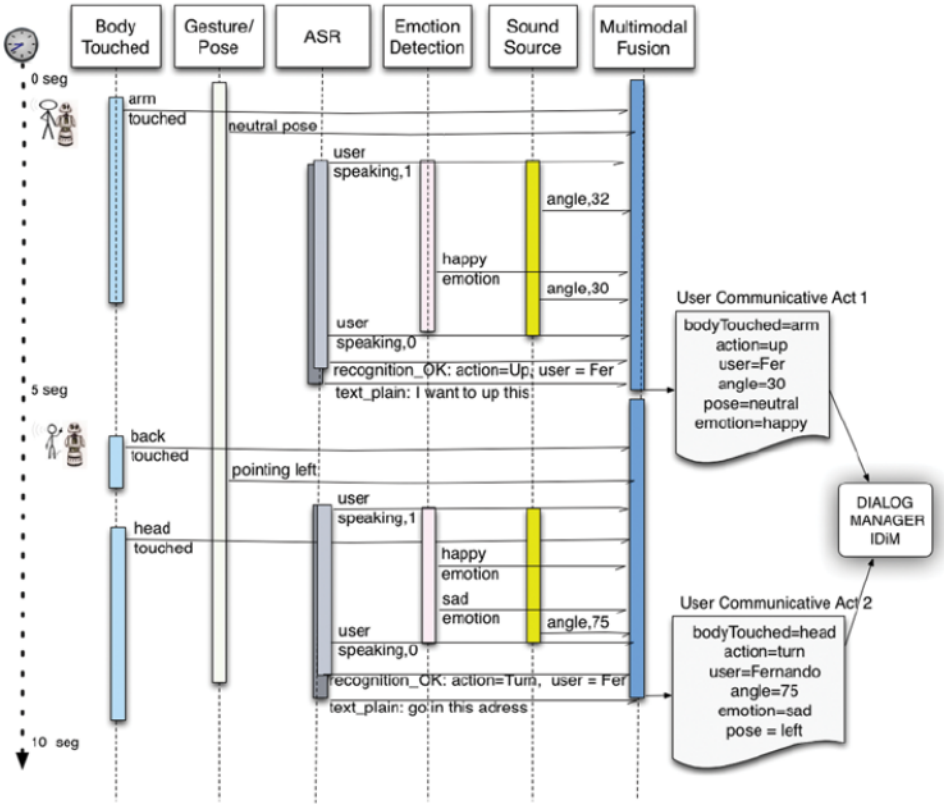


**FIGURE 7** Examples of multimodal fusion as communicative acts (color figure available online).

One of the main advantages of multimodality is related to the possibility of solving deictic references (deixis) using the information given by several input channels due to the complementarity of their information. Deictic references are references to things that are not present in the linguistic context. In this sense, in a deictic expression such as "go there," pointing in a direction, "there" is solved through the component that gives the dialogue in the pose of the user (the pose of his or her body pointing at a particular place). In another example of deixis, the user says "up" while touching one of the robot's arms at the same time. In this case, the part of the body involved in the interaction is given by the module related to the touch.

## SCENARIOS OF THE MULTIMODAL INTERACTION

In this section the main scenarios where the RDS is used are described, analyzing the multimodal fusion process.

The *first* scenario corresponds to a master–slave kind of interaction; therefore, the user commands the robot through the dialogue, by activating and deactivating robot skills. In this scenario, all of the inputs and outputs of information are used and fused/fissioned.

The *second* scenario allows the user to control the robot by voice, and it is specifically designed to test the multimodal fusion (the output of the system is not multimodal) to solve deictic references.

The *third* and *fourth* scenarios present two novel systems, simple but very powerful: the music player and the open cloud dialogue, respectively. In both scenarios the fusion of information is made through two different voice recognizers working in parallel.

## First Scenario: Action Integrator Dialogue

During the HRI the user can order the robot to execute some tasks. The multimodal dialogue system developed acts as an interface between the user and Maggie's skills repertory. Therefore, the RDS is an integrator of the robot skills. As already mentioned, during this kind of interaction the robot plays a slave role, obeying the user's commands.

The dialogue starts when the robot greets the user, by voice and other sounds. Then, the user must greet the robot back, typically by voice, although other channels can also be used during the rest of the interaction, such as RFID tags. When the interaction is made by voice, if the robot recognizes the user's voice tone, then it charges the user profile automatically in order to better adapt the dialogue to its interlocutor (language, experience level, emotional state, preferences, etc). On the other hand, if the robot does not recognize the user, it asks for his or her registration. In the affirmative case, the user must answer some questions about his or her name, age, preferred language, etc. In addition, during this process the robot learns the user's voiceprints for future identifications.

Once the user has been identified, he or she can order the robot to execute any available skill. Currently, the repertory is extent and includes skills such as medicine detection, TV control, to follow the user, to dance, to play games, etc. Depending on the selected skill, the interaction can be carried out in different ways: by voice, using vision, by gestures, etc. Therefore, in this case, all of the multimodal possibilities are exploited.

## Second Scenario: Teleoperation Dialogue

Focusing on the multimodality of the RDS, which allows it to solve deictic references, we have designed a dialog to command the robot in order to evaluate the multimodal fusion. During the interaction, the user must indicate the action to execute using any of the available input channels (voice, touch, vision, and gestures), see Figure 8.

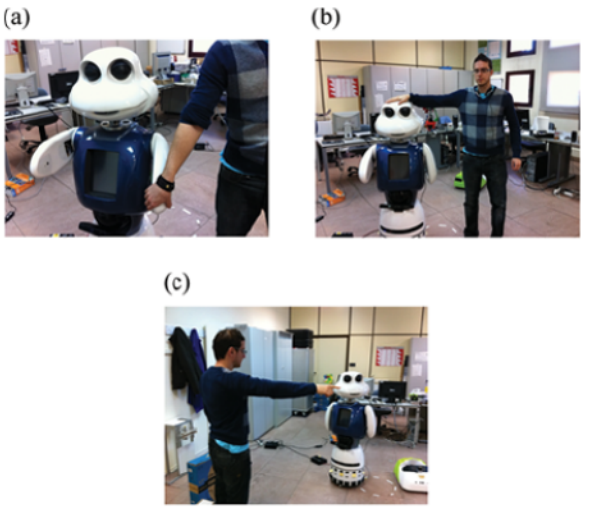


**FIGURE 8** Examples of deictic teleoperation dialogue (a) raise the right arm; (b) turn the head right; (c) go there (color figure available online).

As already stated, the dialogue in charge of this teleoperation is specified in an XML file. In this dialogue, there are some slots of information that must be filled: the action to execute, the part of the body, and action direction. These slots can be filled using several input channels of the RDS.

Let us show some valid examples:

- The user says "raise the right arm" but does not touch any part of the robot's body or execute any bodily gesture. In this case, all of the information needed is present in the sentence. Therefore, *action = raise, body part = right arm*.
- The user says "raise" and, at the same time, touches the robot's right arm. It is necessary for the user to touch the robot's arm before the voice recognition finishes. In this case, *action = raise, body part = right arm*: see Figure 8a.
- The user says "turn right" and, at the same time, touches the robot's head. In this case the user wants the robot to turn its head right, so *action = turn right, body part = head*: see Figure 8b.
- The user says "turn your head left" without touching any part of the robot. In this case, there is no deixis, because all the information is contained in the sentence. Therefore, *action = turn left, body part = head*.
- The user says "go in that direction" while pointing in the desired direction with his or her arm. In this case, the pose of the user indicates the movement direction, so *action = move left, body part = base*: see Figure 8c.
- The user says "go to the right" without touching or indicating anything else with the body. Again, in this case the information needed is present in the sentence. Therefore, *action = move right, body part = base*.

- The user says "go to the charge station" without touching or making gestures. In this case, the robot will move to the position of the charge station, since all the information needed is given in the sentence, so *action = move charge station, body part = base.*
- The user says "go to the door" and nothing else; so again, the sentence gives all of the information: *action = move door, body part = base.*

In Figure 9 a diagram of the conceptual structure of the dialogue presented in this scenario is shown. The rest of the inputs received by the IDiM, such as the sound source localization, user emotion, arousal level, etc., are not considered in this particular case. The multimodal outputs are the gestures involved in the intention of the CA: to move around, to move a limb up or down, to turn the head left or right and the voice to communicate the action that the robot is going to execute.

### Third Scenario: Cloud Music Player

Another scenario that illustrates the multimodal possibilities of the presented dialogue system is one that allows the robot to reproduce the majority of the songs of any music group that can be found on the Internet. The user only needs to say the name of the song or group, or both, and the robot, almost instantly, begins to reproduce the requested song.

In this case, the multimodality is present at the entrance of the system, in which two available recognizer engines are used. One of them is based on grammar (Loquendo ASR), and the other is based on open-grammar or plain text (the Google ASR web service), and their functionality and use are different. In addition, another output of the system is incorporated: the music player.

In order to implement them, the musical player and interchange service Goear[14] have been used, as well as a script in Perl language[15], which facilitates the reproduction of one or several songs of a band from the line command. This allows the dialogue system to execute the script by passing the name of the song or group as arguments.

The dialogue specified in this scenario needs to fill two slots: action and name.

In this scenario both recognizers are fused. When the user tells to the system that he or she wants to hear music, the system fills a slot of information based on grammar (using Loquendo ASR), which corresponds to the action "play." Next, the dialogue uses another form where the slot to be filled is plain text, and it serves to know the song or group to be played. This plain text is filled by the Google ASR recognizer. This recognizer does

[14]http://www.goear.com
[15]http://www.splitcc.net/Downloads/playgoear.pl

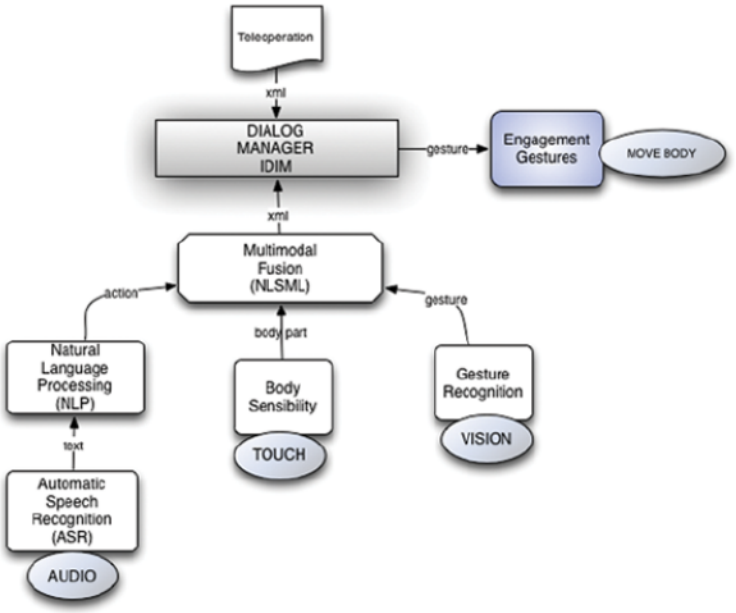


**FIGURE 9** Deictic teleoperation by dialogue (color figure available online).

not generate semantic results: it rather gives back the textual transcription of the sentence (given by the user). The other solution—that is, the use of Loquendo to fill the slot corresponding to the name of the song or group—is not practical because it would imply having very large grammar containing all possible songs and groups.

### Fourth Scenario: Open Cloud Dialogue

In this last scenario the robot is able to answer almost any question that the user can ask verbally by voice, and in any language. For example:

- "What is the distance between the Earth and the moon?"
- "Where am I?"
- "What is the name of the capital of France?"
- "What do you know about Rafa Nadal?"
- "What is the whether in New York?
- "How much is $12 \times 3$?"

In this case, we have designed a dialogue system very similar to the well-known Siri,[16] which is able to answer complex questions.

[16] http://www.apple.com/iphone/features/siri.html

Firstly, the user must tell the system that he or she wants to ask a question: then he or she asks the question. The robot repeats the question to the user and, a few seconds later, it answers by voice.

The dialogue system works in a way very similar to that in the previous scenario. First, it fills the slot about the action information using Loquendo ASR and then moves to the next form where the slot about the question is plain text. This slot is filled using the Google ASR recognizer, which is able to translate the question into many languages.

If the user is interacting with the robot in a non-English language, the system uses the Google or Microsoft on-line translators to translate the question into English. This translated question is sent to the semantic metha-searcher Wolfram Alpha,[17] which answers the question with the available information in its knowledge database, using an XML file.

Once the answer is obtained, the XML file is parsed to get the most interesting parts of the answer (mainly, those that can be communicated by voice). Then, they are synthesized by the dialogue system in the original language, for which use of the translator is necessary again. If the translations are not correct, some important information can be lost. Moreover, it may happen that there are certain questions to which the system is not able to give an answer. In those cases, the user is informed that there is no any available information about the question.

Another aspect that has been considered in order to make the system more interactive is the velocity of response of the dialogue. For this reason, a detection system has been implemented at a low level to detect the beginning and the end of the voice. Therefore, the voice samples are sent in a compressed format (flac or speed) to the Google ASR service as soon as they are available. Therefore, the translation is obtained with a one-second of delay since the phrase is ended. At the same time, the Wolfram Alpha service and the parsing of the answers take about 3 to 4 s, although this time is used by the robot to repeat the question.

## CONCLUSIONS AND FUTURE WORKS

In this work, RDS in which multimodality is one of the basic features is presented. Based on the communicative acts (CAs) theory among humans, its use in the multimodal HRI is introduced. The information received from the different input channels is fused in order to create a CA, which represents the message to be transmitted by the user during each turn of dialogue. In order to explain the multimodal fusion process, four different scenarios are presented where the dialogue system can exploit its multimodality at both

[17]http://www.wolframalpha.com/

the input and output channels by executing different tasks and dialogues in a natural and successful way.

As future work, it would be desirable for the dialogue system to be able to solve sentences such as "go to that place" or "go there" without touching but by making a gesture pointing to the place where the user wants the robot to move. In this case, some extra functionalities of the robot are needed. For example, to know the exact position of the user on the map, to recognize the direction in which the user is pointing, etc.

## ACKNOWLEDGMENTS

The authors gratefully acknowledge the funds provided by the Spanish Government through the project "A New Approach to Social Robotics (AROS)" by the Ministry of Science and Innovation. The research leading to these results has received funding from the RoboCity2030-II-CM project (S2009/DPI-1559), funded by Programas de Actividades I+D en la Comunidad de Madrid and cofunded by Structural Funds of the EU.

## REFERENCES

Alonso-Martin, F., A. A. Ramey, and M. A. Salichs. "Maggie: El Robot Traductor." In *Workshop RoboCity2030 II*, 57–73. RoboCity, 2011.

Bach, K. and R. Harnish. *Linguistic Communication and Speech Acts*. 1982.

Barber, R. and M. Salichs. "A New Human Based Architecture for Intelligent Autonomous Robots." In *Intelligent Autonomous Vehicles 2001 (IAV 2001): A Proceedings Volume from the 4th IFAC Symposium, Sapporo, Japan, 5–7 September 2001*, 81. Pergamon, 2002.

Bennett, C., A. Llitjós, S. Shriver, A. Rudnicky, and A. W. Black. "Building VoiceXML-Based Applications." In *Seventh International Conference on Spoken Language Processing*, 2–5. Citeseer, 2002.

Black, A. W., S. Burger, B. Langner, G. Parent, and M. Eskenazi. "Spoken Dialog Challenge 2010." In *2010 IEEE Spoken Language Technology Workshop*, 448–53. Berkeley, CA: IEEE, 2010.

Bohus, D. and A. I. Rudnicky. "The Raven-Claw Dialog Management Framework: Architecture and Systems." *Computer Speech & Language* 23, no. 3 (2009): 332–361.

Bruner, J. "The Ontogenesis of Speech Acts." *Journal of Child Language* 2, no. 1 (1975): 1–19.

Cassimatis, N., J. Trafton, M. Bugajska, and A. Schultz. "Integrating Cognition, Perception and Action through Mental Simulation in Robots." Naval Research Laboratory 2004. Technical report.

Cheyer, A., L. Julia, H. Bunt, R.-J. Beun, and T. Borghuis. *Multimodal Human–Computer Communication,* Volume 1374 of *Lecture Notes in Computer Science*. Berlin, Germany: Springer, 1998.

Corrales, A. and M. A. Salichs. "Integration of a RFID System in a Social Robot." *Computer and Information Science* 44 (2009): 63–73.

Dominey, P. F., A. Mallet, and E. Yoshida. "Progress in Programming the hrp-2 Humanoid Using Spoken Language." In *International Conference on Robotics and Automation (ICRA07)*, 2007.

Eberman, B., J. Carter, D. Meyer, and D. Goddeau. "Building VoiceXML Browsers with Open VXI." In *Proceedings of the Eleventh International Conference on World Wide Web WWW'02*, 713. New York: ACM Press, 2002.

Falb, J., R. Popp, T. Rock, H. Jelinek, E. Arnautovic, and H. Kaindl. "Fully-Automatic Generation of User Interfaces for Multiple Devices from a High-Level Model Based on Communicative Acts." In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. 26–26. IEEE, 2007.

Fry, J., H. Asoh, and T. Matsui. "Natural Dialogue with the Jijo-2 Office Robot." In *1998 IEEE/RSJ International Conference on Intelligent Robots and Systems, Proceedings*, 1278–283. 1998.

Gorostiza, J., R. Barber, A. Khamis, M. Malfaz, R. Pacheco, R. Rivas, A. Corrales, E. Delgado, and M. Salichs. "Multimodal Human–Robot Interaction Framework for a Personal Robot." In *ROMAN 2006 The 15th IEEE International Sym posium on Robot and Human Interactive Communication*, 39–44. Hatfield, UK: IEEE, 2006a.

Gorostiza, J., R. Barber, A. Khamis, M. Malfaz, R. Pacheco, R. Rivas, A. Corrales, E. Delgado, and M. Salichs. *Multimodal Human–Robot Interaction Framework for a Personal Robot*. IEEE, 2006b.

Gorostiza, J. F. and M. A. Salichs. "Natural Programming of a Social Robot by Dialogs." In *AAI 2010 Fall Symposium*, 20–25. 2010.

Hüwel, S., B. Wrede, and G. Sagerer. "Robust Speech Understanding for Multi-Modal Human-Robot Communication." In *15th IEEE International Symposium on Robot and Human Interactive Communication (RO MAN06)*, 45–50, 2006.

Iba, S., C. J. Paredis, and P. K. Khosla. "Interactive Multimodal Robot Programming." In *2002 IEEE International Conference on Robotics and Automation*. IEEE/RSJ, 2002.

Kibria, S. and T. Hellström. "Voice User Interface in Robotics Common Issues and Problems." 2007. Available at: http://aass.oru.se.

Larsson, S. and D. R. Tram "Information State and Dialogue Management in the TRINDI Dialogue Move Engine Toolkit." *Natural Language Engineering* 6, nos. 3–4 (2000): 323–40.

Lemon, O., A. Gruenstein, A. Battle, and S. Peters. "Multi-Tasking and Collaborative Activities in Dialogue Systems." In *Proceedings of the 3rd SIGdial workshop on Discourse and Dialogue Volume 2, SIGDIAL '02*, 113–24. Stroudsburg, PA: Association for Computational Linguistics, 2002.

Lucas, B. "VoiceXML for Web-Based Distributed Conversational Applications." *Communications of the Association for Computing Machinery (ACM)* 43, no. 9 (2000): 53–57.

Nigay, L. and J. Coutaz. "A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion." *Proceedings of the INTERACT'93 and CHI'93*, 1993.

Niklfeld, G. and R. Finan. "Architecture for Adaptive Multimodal Dialog Systems Based on VoiceXML." In *Proceedings of EuroSpeech*, 1–4. Association for Computational Linguistics, 2001.

Perzanowski, D., A. Schultz, W. Adams, E. Marsh, and M. Bugajska. "Building a Multimodal Human-Robot Interface." *IEEE Intelligent Systems* 16, no. 1 (2001): 16–21.

Reithinger, N. and J. Alexandersson. "SmartKom: Adaptive and Flexible Multimodal Access to Multiple Applications." In *ICMI '03 Proceedings of the 5th Interna tional Conference on Multimodal Interfaces*. 2003.

Salichs, M., R. Barber, A. Khamis, M. Malfaz, J. Gorostiza, R. Pacheco, R. Rivas, A. Corrales, E. Delgado, and D. Garcia. *Maggie: A Robotic Platform for Human Robot Social Interaction*. IEEE, 2006.

Searle, J. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

Searle, J. "Indirect Speech Acts." *Syntax and Semantics* 3 (1975): 59–82.

Seneff, S., D. Goddeau, C. Pao, and J. Polifroni. "Multimodal Discourse Modelling in a Multi-User Multi-Domain Environment." In *Proceedings of the Fourth International Conference on Spoken Language Processing. ICSLP '96*, Volume 1, 192–95. IEEE.

Shimokawa, T. and T. Sawaragi. "Acquiring Communicative Motor Acts of Social Robot Using Interactive Evolutionary Computation." In *2001 IEEE Interna tional Conference on Systems, Man and Cybernetics. e Systems and e Man for Cybernetics in Cyberspace*, Volume 3, 1396–1401. IEEE, 2001.

Stiefelhagen, R. "Natural Human-Robot Interaction Using Speech, Head Pose and Gestures.". In *Intelligent Robots and Systems*, Volume 3, 2422–227, 2004.

Stiefelhagen, R. and H. Ekenel "Enabling Multimodal HumanRobot Interaction for the Karlsruhe Humanoid Robot." *IEEE Transaction on Robotics* 23, 840–851, 2007.

Toptsis, I., S. Li, B. Wrede, and G. A. Fink. "A Multi-Modal Dialog System for a Mobile Robot." In *International Conference on Spoken Language Processing* Volume 1, 273–76. 2004.

Wahlster, W. "Smartkom: Symmetric multimodality in an adaptive and reusable dialogue shell." *Proceedings of the Human Computer Interaction Status*. 2003.

Waibel, A. and B. Suhm. "Multimodal Interfaces for Multimedia Information Agents." In *1997 IEEE International Conference on Acoustics, Speech, and Signal Proces sing*, Volume 1, 167–70. IEEE Computer Societ Press, 1997.

Walker, M. A., R. Passonneau, and J. E. Boland. "Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems." In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics ACL '01*, 515–22. Morristown, NJ: Association for Computational Linguistics, 2001.

Yin, Y. "Human–Humanoid Robot Interaction System Based on Spoken Dialogue and Vision." In *2010 3rd International Conference on Computer Science and Information Technology*, 328–332. IEEE, 2010.