Voice User Interfaces for Service Robots: Design Principles and Methodology

Pepi Stavropoulou[1], Dimitris Spiliotopoulos[2],
and Georgios Kouroupetroglou[1](✉)

[1] National and Kapodistrian University of Athens, Athens, Greece
{pepis,koupe}@di.uoa.gr
[2] University of the Peloponnese, Tripoli, Greece
dspiliot@uop.gr

Abstract. This work presents the concerns, prerequisites, and methods for building interaction interfaces for service robots. It mainly deals with Voice User Interfaces - VUI (also called Spoken Dialogue Interfaces - SDIs) but also includes issues on multimodal interfaces, involving speech and other modalities. Human-machine interaction in the area of robotics raises certain challenges that respective interface design for other domains ignores. Robots, and more importantly, service robots, execute actual tasks based on plans and scenarios that, in effect, layout their usage. The completion requirements, as well as the workflow needed for those tasks, form a very significant set of rules that affect and sometimes govern the interaction between the user and the machine. Those rules are embedded to the design of the interaction system and, together with the communicated context, provide the sets and constraints that the system is based upon. These constraints can be realized in the form of specific dialogue management design, dialogue flow, belief states models, verification, disambiguation, and grounding techniques as well as more subtly use of specific speech and dialogue acts – all the above affect all stages of the lifecycle. Moreover, significant merit goes to usability, and the techniques for its evaluation, issues that are of the utmost importance when any user-machine interface is designed and assessed.

Keywords: Voice User Interfaces · Service robots · Spoken dialogue interaction · Usability evaluation · Computer-mediated communication

1 Introduction

In the recent years, the technology has allowed the idea of robotic assistants and services to become feasible for certain domains (health-related assistance, disability, help for the elderly, office assistants, etc.), where they are used for a variety of tasks by a variety of non-expert users. Nowadays, service robots have been developed to assist people at their homes and workplaces, performing household chores, delivering objects and responding to inquiries about the weather or the TV program among others. In this sense, it has become increasingly important that robots are designed to become part of the lives of ordinary people, enabling a more natural, intuitive, and effective mode of communication. To address this requirement, a considerable amount of research has

M. Antona and C. Stephanidis (Eds.): HCII 2020, LNCS 12188, pp. 489–505, 2020.
https://doi.org/10.1007/978-3-030-49282-3_35

been dedicated to the development of Spoken Dialogue Interfaces (SDIs) as a cornerstone aspect of Human-Robot Interaction (HRI). This trend has become more obvious now as a key area of research, spanning from mobile devices to robotics.

Voice User Interfaces utilize spoken language as the most natural and powerful means of human communication to maximize the usability of human-machine interfaces. Depending on the particular type of SDIs, different levels of flexibility and robustness in handling spoken input and output are allowed. Thus, Spoken Dialogue Systems (SDSs) may range from simple finite-state systems that handle a limited number of commands, to more advanced systems capable of inference and planning as part of a more collaborative view of interaction.

While in general-purpose SDSs, such as call routing or ticketing, the term VUI has mostly been reserved for systems that use open-ended prompts, large vocabularies, and flexible dialogue structure, for HRI in particular, the term has sometimes been used to describe systems that employ speech in the simple form of a command language as opposed or in addition to other less intuitive modalities such as GUIs and buttons. Such systems, however, are far from resembling natural human-human conversation, and their deviation from the user's natural discourse patterns often places a considerable load on cognition and hinders learnability On the other hand, the collaborative and socially-oriented nature of a service robot's tasks has led to the development of architectures that can incorporate advanced natural language processing techniques involving inference, dialogue act identification and anaphora resolution among others; more advanced practical systems have started to emerge, which – even though they make limited use of understood vocabulary and syntax – are an important first step towards truly natural HRI.

In the following sections, we will first present the basic implications, advantages, and disadvantages, as well as motivation for building SDIs for service robots. Next we will elaborate on the communicative principles that are especially important for the development of usable systems and as such map outline the requirements for the interface design. Sections 4 and 5 discuss design and evaluation methodologies for service robots. Finally, Sect. 6 touches upon more advanced subjects such as social skills, emotions, autonomy, and adaptation in robotic systems.

2 Motivation for Spoken Dialogue Interfaces

Simply put, SDI is the means of communication that the majority of people are inherently equipped and familiarized with since birth. It is the most intuitive, natural, and powerful tool in human communication. As such, it constitutes the most promising means of interaction in the emerging field of service robots, in which nonexpert users need not just to operate the robot but work with it in order to accomplish specific tasks. This is particularly important if one considers that a significant application area for service robots is helping the elderly, who are most often "technology-unaware" and have difficulties interacting in ways they are not familiar with [1, 2]. In this respect, spoken language interfaces that can only handle a limited set of commands with a fixed set of arguments, rigid preconditions, and task completion requirements (cf. points on "under-specification" in the following section) are similarly ineffective and unfriendly, as the user needs to memorize these commands adjusting to the interface contrary to a more user-centered approach in which the interface adapts to the user. Likewise, companion robots – robots that engage in conversations in a socially acceptable manner, often displaying personality traits and emotions – constitute systems that, by definition, go beyond the realm of simple command and control interfaces.

Furthermore, as service robots evolve, the variety and number of tasks they can perform increases, while at the same time, they are involved in a constantly changing, dynamic situation setting, which makes it difficult to pre-specify tasks at hand and model them in the form of stand-alone commands. The latter necessitates the use of spoken language for the development of an easier to use interface that compensates for the complexity of tasks implemented in the task manager, by providing a more flexible dialogue structure, loosening requirements on user input, handling under-specification, reference, feedback, and grounding. In addition, VUIs allows users to teach the robot contributing to an increase in robot's adaptation, learning ability and autonomy [3, 4]. In general, there is a tendency that service robots are no longer considered as simple tools, means to an end but as collaborative partners with specific social and communicative skills [5]. This view is served better through spoken language and is often intensified by the physical embodiment of the robot itself. Humanoid appearance especially triggers certain expectations on the part of the users that are, in turn, more likely to apply human-human communication principles for HRI compared to ordinary human-computer interfaces.

Finally, the use of speech ensures a design-for-all approach to robotic system design. Universal Design and accessibility denote that an application is designed in such way so that it may be used by people "with different abilities, requirements and preferences in a variety of contexts of use" [6]. Apart from providing an alternative means of interaction to users such as the elderly or people with limited dexterity, speech is also best suited for hands/eyes busy situations and multitasking. The latter is common for home or office users that may be engaged in other tasks while addressing the service robot. In the same manner, multimodality is important, enabling robust communication in situations where speech is not optimal (e.g., high noise levels, workgroup settings, teleoperation). After all, natural human conversation is not restricted to speech but is accompanied by other means such as gestures, facial expressions or posture that also convey attitudes and meaning.

The above claims are corroborated by surveys conducted to assess users' preference of modalities, in which users demonstrated a preference for spoken language as a means of interaction. Torta et al. [7] report a clear user preference for natural spoken language, followed by touch screen, gestures and command language, when interacting with a household service robot [7]. Still, overall results indicated that users favor the availability of more than one, complementary means of interaction, opting for multimodality, where available. It is still the case that for specific service robot interfaces, the detection of the presence of the user as well as the activity can be recognized and modeled in such ways as to trigger multimodality [5].

On the other hand, there are certain drawbacks in using VUIs, which – if not taken into account – may deprive the system of any practical use. First of all, speech recognition conditions in real-life settings may involve high levels of noise-causing degradation of the recognizer's performance. To address automatic speech recognition

(ASR) limitations, most robotic systems use a limited vocabulary set reaching a few hundred words at most [1, 8]. By limiting the recognizer's search space, they could achieve over 90% recognition success rate under lab test, normal/low noise conditions [9]. As a drawback the out of vocabulary rate (OOV) in users' utterances addressed to the system may increase especially in non-controlled, real-life conditions with little or no user training. Other issues concern the difficulty of long-distance (far-field) speech recognition, identification of the voice source, and identification of commands addressed to the robot in workgroup environments.

Finally, another important parameter that should be taken into account is the degree of computational efficiency affected by the use of sophisticated and computationally costly speech processing algorithms. It should be noted that even without SDI capabilities, service robots can be very complex, comprised of several modules working in parallel (e.g., navigation, visual object identification, task planning) that must respond efficiently in real-time. This necessitates VUI techniques that are easy to specify and maintain and lead to robust and fast input processing.

3 Related Work

A typical spoken dialogue system embedded in a service robot consists of the following components:

- The Speech Recognizer that converts the user's spoken input into a text string. Typical speech recognizers for service robots handle only a limited in- domain vocabulary. Hand-written, context-free rule-based grammars are used that determine the recognizer's space based on the developer's expectations of what users are likely to say [10]. Alternatively, some works have utilized real use corpora collected through Wizard of Oz settings and user testing [1] or handwritten utterances based on usage scenarios [8], in order to train statistical language models for recognition.
- The Natural Language Understanding (NLU) module that semantically interprets the string passed by the speech recognizer. A commonly used method for semantic analysis is based on semantic augmentations attached to grammar and lexicon rules to fill in dedicated slot values or construct predicate-like meaning representations relative to the task at hand for service robots [11].
- The Dialogue Manager (DM) that evaluates and reassess the NLU input with regards to dialogue history, conversation principles, specific task, domain and user knowledge, in order to decide upon the next dialogue step and fulfill a specific strategy.
- The Natural Language Generation (NLG) component produces an appropriate concrete language response based on the DMs abstract input. Simple NLG techniques that are most commonly used involve template filling methods, in which system utterances are, to a large extent, predefined. Otherwise, more advanced methods involving discourse planning and surface realization of utterances may be used [12]. Also, some NLG components add prosodic annotations to the word string, providing an enriched input to Concept-to-Speech synthesizers [13, 14].

- The Speech Synthesizer that converts NLG text input to speech. Typically, off the shelf speech synthesizers are used that are naive to discourse structure and context properties. Concept to Speech synthesizers, on the other hand, are more advanced systems that may utilize contextual information passed from the NLG module for producing appropriate, context-aware utterance intonation [15, 16].

Depending on the technology used and the specifics of the dialogue management component, in particular, there are three main types of spoken dialogue systems [12, 17]: state-based, frame-based, and more advanced information state architectures. State-based and frame-based systems have been the most commonly used so far due to their ease in development and low computational cost.

State-based systems represent dialogue as a predefined series of states, whereas the user is expected to provide specific input in a particular order. This makes the user's utterances easier to predict, leading to faster development and more robust recognition and interpretation at the expense of limited flexibility in the structure of the dialogues. Their limited flexibility, however, often renders them less appropriate for complex tasks. Frame-based systems represent dialogue as a list of slots each slot corresponding to pieces of information that the system needs to acquire to perform a task. They offer a higher level of flexibility compared to state-based models, as the dialogue is not completely pre-determined, and a limited level of mixed-initiative is allowed. That is, the system formulates questions to fill in particular slots that remain empty, but the user may take the initiative in the dialogue and provide more information than asked. This additional information is used to fill in more slots, saving the user from having to answer subsequent questions, and leading to more efficient dialogues compared to state-based approaches [9, 18–20]. Some of these approaches are combined with more advanced features such as pronoun resolution or basic speech act identification, increasing the system's robustness while minimizing computational cost [10].

Information state systems, on the other hand, make use of sophisticated discourse models in order to represent and update dialogue context, interpret and generate dialogue acts, identify, form, and execute conversational goals and plans. Such systems are equipped with advanced inference, reference resolution, speech act interpretation and grounding capabilities. As such, they can accommodate a greater degree of flexibility and mixed-initiative and are suitable for complex, collaborative tasks where the series of actions that need to be performed and the particular pieces of information required are hard to predict in advance. Their implementation and maintenance, however, is far more complex and computationally expensive compared to state or frame-based systems. Wilske and Kruijff (2006) presented an example of a service robot that incorporates a more advanced, information state architecture [11]. The system uses a BDI (Belief, Desire, Intention) module to mediate between subsystems for different modalities. It exploits knowledge about the preceding discourse, the situational context, and the task in order to referentially and rhetorically resolve the current utterance's linguistic context, infer user goals through indirect speech act identification, take initiative, ask for help and clarifications when necessary, and learn about the environment it operates in through the understanding and production of natural language.

4 Service Robot HRI: Communication and Design Principles

Service robots situated in people's everyday lives aiming to co-operatively accomplish specific tasks, often serving as human companions, need to interact with people on a more social level. It is, in fact, this elevated, enriched form of interaction with people in natural, unstructured, everyday environments that fundamentally differentiates them from traditional industrial robots. Though limitations of current technology render a truly natural, human-like HRI an issue of a not so near future, HRI design could benefit from incorporating knowledge of human communication principles. Taking into account people's well known tendency to attribute human-like characteristics to machines [21] it is reasonable to expect that people will be inclined to apply human-like conversation principles especially when interacting with a robot whose physical stature may encourage such behaviour. In fact, it is no wonder that people unconsciously apply conversational behavior that is implicitly learned and used from a very young age even when they are advised against it. Hüttenrauch et al. (2003) reported that people used gestures to navigate a service robot even though they were told beforehand that the robot was incapable of understanding such input [22].

At the heart of each dialogue, determining conversational behavior is the communicative situation itself. The "who", "when", "where", "why" and "what about" of communication determine the form and the content of the message. An example application of this principle is user modeling. Robots designed as museum tour guides, for instance, utilize knowledge of the humans that they will interact with (adults, children, experts, artists), in order to properly adjust their personality, behavior, and roles.

Most importantly, though, understanding of the situational context is a prerequisite for effective interaction and successful task fulfillment. A crucial difference between service robots employing spoken natural language and other spoken dialogue systems is the importance of the situational – including the visual – context for the former. Human-robot dialogue is a principally situated dialogue, "a spatially embedded interaction" [23, 24] in the sense that robots need to identify and perform actions on elements of a shared environment having established a correspondence between the human and the robot's perception of the environment's spatial organization. To do that, robots need to make and resolve reference to temporal and spatial aspects of the interaction, interpret pronouns, ellipsis and so forth. Such requirements lead to the adoption of more advanced dialogue management techniques and discourse models that make use of rich dialogue history and context representation, as well as sophisticated inference mechanisms based on task knowledge, knowledge of conversation principles, and current information state in general. Seemingly simple commands such as "Turn right here" or "Bring it to me" involve the not so trivial task of resolving anaphoric expressions such as "here" and "it" to salient discourse referents.

Furthermore, as users cannot be expected to unambiguously provide all information required for the robot to perform an action, robotic systems further need to address under-specification, incomplete user input, which does not fulfill the robot's knowledge preconditions. The omission of some detail is almost inevitable in all human communication. In an experiment examining spatial, direction tasks using a wheelchair robot, Tenbrink and Hui (2007) reported that users were often vague in their descriptions as well as unaware or uncertain about the level of detail that is required for the robot to unambiguously establish a spatial goal [23]. Therefore, advanced dialogue modeling techniques were required in order to either infer missing information based on discourse context or explicitly ask for it through clarifications and info requests. With regard to the latter, Tenbrink and Hui [23] point out that clarifications should depend on discourse history and be formulated based on previous user input and grounded knowledge rather than being generic clarifications [23]. This way user's and robot's perceptions are better matched, and confusion and uncertainty are reduced.

Another pertinent and most significant aspect of communication is grounding [25, 26]. Grounding is the establishment of common ground among the interlocutors. The term refers to the goal and process of achieving mutual understanding within the dialogue and acknowledging this understanding, thus making the other participant confident of the progress made to fulfill the dialogue's goal. The establishment and communication of shared understanding are primarily achieved through feedback. There are several means for providing feedback, both verbal and non-verbal. Examples of the former are relative next turns, verbatim repetition or paraphrasing of the interlocutor's previous utterance, backchannels such "uh-huh" and "hmm," explicit acknowledgments such as "I see," use of discourse markers such as "well" or even emotional prosody providing feedback on speaker's attitude. Non-verbal means for production of feedback, demonstration of attention and awareness are eye-gaze and face/object tracking mechanisms, as well as simple gestures such as nodding or pointing. Even a blinking indicator light on the robot may provide feedback that the system is on and hearing. Building on Clark and Schaefer (1989) [26], Brennan and Hulteen (1995) proposed a multimodal model of eight levels of feedback associated with specific system states, ranging from pointing out that the system is attending or not to notifications regarding intent and initialization of task execution, and reporting on task execution outcome [27]. This model was partly implemented in the development of Cero, a mobile service robot for object delivery [5].

A widely used strategy for achieving common ground, providing feedback and addressing potential problems in understanding is confirmations and clarifications. Confirmations may be explicit or implicit. In the former case, the system directly assesses the correctness of its understanding by asking a targeted yes/no question. In the case of implicit confirmation, the system combines what has been understood with a question for a missing argument in a theme-rheme informational organization of the produced utterance [23, 28, 29]. Note that, again, verbal confirmation may be accompanied and reinforced by appropriate gestures. The confirmation strategy followed – explicit or implicit – depends on various parameters: ASR confidence scores and error cost estimation are most commonly used [30], while robotic systems may also use task knowledge and dialogue history to identify inconsistent commands or plan execution failure and decide upon the subsequent dialogue act (e.g., confirmation, elaboration, clarification etc.). For example, when a robot recognizes that it cannot fulfill a request (e.g. the user asks the robot to fetch an object that is not part of the shared spatial organization), it may ask for confirmation or clarification.

Studies with service robots have demonstrated the importance of feedback for the quality and efficiency of the interaction [31]. Observations have been reported with regards to users of standard SDSs who are often confused when the system does not explicitly acknowledge shared understanding [32]. In general, grounding and feedback are especially important for HRI, also given the limitations of current ASR and NLU systems as well as users' proclaimed skepticism and occasional lack of trust towards new generation robotic systems.

Another aspect of communication that is often exploited by current robotic systems is the notion of speech dialogue acts. There are three types of speech acts [33]: (a) locutionary acts, that is the utterance that is produced and its literal meaning, (b) illocutionary, the acts that the speaker performs when producing this utterance, e.g., asking, asserting, requesting, etc. and (c) perlocutionary acts, the result of the utterance upon the hearer's beliefs, actions and so forth. A robotic system should be able to identify and reason about speech acts, in order to identify the user's intentions and plan its course of actions accordingly. However, identification of speech acts is not a trivial task, as they are based on the speaker's cognitive state, and there is no one to one correspondence between surface syntactic structure and illocutionary act type. For example, a sentence such as "Can you bring me a cup of coffee?" could in principle be a yes/no question or a request. Therefore, systems that merely make use of syntactic mood to identify speech acts risk misinterpreting users' intentions even for small domains. For an example of a more advanced BDI model that infers illocutionary and perlocutionary acts based on plan recognition techniques, interested readers may refer to Allen (1995) [34]. Wilske and Kruijff (2006) also present a more sophisticated approach to indirect speech act identification for service robots; that is identification of illocutionary acts that are produced with a syntactic form other than the one they are conventionally associated with (for example an interrogative utterance that is used to perform a request instead of an imperative) [11].

5 Designing Service Robot HRI

There are five main stages in the lifecycle of a Spoken Dialogue Interface:

- Requirements specification and initial planning
- Design
- Implementation and testing. The SDI components are developed and integrated with other system components. Unit, system and user testing is performed
- Deployment. The market-ready system is released to real users
- Evaluation: data is collected from real-life use, and the system is monitored and tuned accordingly

This section focuses on the requirements specification and design steps of the methodology. During these steps, the system functionality is analyzed, and design decisions are made resulting in a complete, detailed specification of the dialogue that serves as input to the development phase. Questionnaires and user interviews, development of usage scenarios [22] are some of the tools employed at this early stage. Furthermore, WOZ simulations [35] are the dominant method for evaluating early design choices for service robots SDIs [22, 23, 36].

In order to decide on key dialogue characteristics, designers need to perform thorough analyses of the users, their goals and needs, of the tasks to be performed, as well as the particular settings, whereas the interaction takes place. The latter is specifically important for the development of service robots, which are particularly sensitive to the situational context. In fact, designers should cater not only for primary user needs but also for bystanders and secondary users that may interfere with robot's task execution [22]. Furthermore, the variability of the scenarios and the spatial organization and context constraints pertaining to a situated interaction place upmost significance on the analysis of the "abstract" communicative situation, in order to maximize system's robustness and usability. In this sense, the development lifecycle should not just be user-centred – and much less robot (system)-centred – but rather usability and situation-centred. That is, the shared world in which the interaction draws the information from, commends the parameters of the communication that are then shaped from the user requirements and the tasks that the service robot is designed to perform. As an example, robots designed as museum tour guides have different knowledge of their environment, the humans that will interact with them (adults, children, experts, artists), their services and roles, and their personality. On the other hand, robotic assistants for the elderly have different requirements, workspace (mostly homes), target users (elderly people), and roles. The requirements for multimodal interaction, noisy environment, multiple users or user groups, personalization (for types of users), and social skills are essential for the former. Robust spoken language interaction, dedicated services for specific needs, simplified interface design, and communication are essential for the latter.

In other respects, standard principles that apply to the development of usable human-machine interfaces, apply to SDIs for robotic systems as well:

- iterative testing, design and build process, whereas design choices are re-evaluated and refined at each iteration
- user involvement from the early stages of the system lifecycle as part of user-centered design
- adherence to conversation principles such as grounding, context awareness and turn-taking
- adherence to general usability principles such as clarity and consistency
- focus on error handling and dialog repair, given that there is no error-free human-machine communication or even human-human communication for that matter
- building on the "natural" mental model that first-time users bring to the interaction, i.e., their existing – and possibly expected – view of the interaction, based on their experience and understanding of how things have worked so far.

The success of an interface greatly depends on the correspondence between this "natural" mental model and the proposed model afforded by the design of the interface [35, 37]. Ideally, a system should build on the users' prior knowledge and experience, in order to create a more familiar, intuitive, easier to learn, user interface. The same principle applies not only to the robot's behavior and language characteristics but to its appearance as well. The Care-O-bot 3 robot [38], for example, contrary to its overall tecnomorphic design, uses a human-like "arm" feature to help users relate to the robot and understand its behavior (when e.g. serving drinks).

Other aspects of interest are the distribution of initiative in dialogue as well as lower-level issues such as signaling the robot's attention. Based on the dialogue initiative strategy employed, systems may range from single to mixed-initiative. In the first case only one participant (system or user) completely controls the dialogue, while in the second case both participants may initiate topics, change the dialogue flow, and adjust their plan in response to the interlocutor's input. Though many current robotic systems are user-directed systems based on a command and control language that minimizes the complexity of the recognition and interpretation process, only mixed-initiative systems can truly serve the view of HRI as collaborative interaction. The following is an example of different sub-goals being initiated by both interlocutors at each dialogue turn, which could be handled by a mixed-initiative system alone. Suppose that the robot is again ambiguously instructed to bring a box; as a result, a clarification question is initiated, such as "Should I bring the red box?". In response, the user may specify the entity at hand based on color or – if for example, the robot has misrecognized the entity's color – use a different attribute such as the object's exact location, e.g. "the box on the table". Now, based on the robot's perception of the environment, there may again be more than one entity that matches this specification. Thus, the robot could either infer the user's goal based on each object's proximity (i.e., if the user is in the kitchen, it is most likely that the referred entity is on the kitchen's table rather than in the living room) or initiate another appropriate clarification request. In general, in mixed-initiative systems, the robot often initiates conversation, goals, and sub-goals, provides suggestions, or may even ask bystanders for help.

On top of being able to address the user and initiate conversation, more significantly, a service robot needs to understand when it is being addressed. This is especially important in workgroup settings where the primary user may be interacting with other people in the robot's proximity, and so system success cannot merely rely on key phrase spotting and recognizer's robustness. Typically, dedicated commands or keywords (e.g. "hello" [10] or "robot") are used to signal the robot's attention. In Baltus et al. (2000) all utterances directed to the robot had to begin with the robot's name "Flo", in order to minimize the probability of responding to utterances not addressed to the robot itself [1]. This, however, brought redundancy to the conversation once it had been initiated. Optimally, systems should make use of other information resources such as face tracking, recognizing face and voice direction and pose, as well as dialogue state and task information in the course of interaction in conjunction with the understood spoken input. The mobile service robot described in Takiguchi et al. (2008) makes use of acoustic features in order to discriminate between commands addressed to the robot and human-human conversations [39]. Other modalities such as on/off buttons and touch screens, may also be used.

Furthermore, with regards to service robots, in particular, physical stature, personality, social and collaborative skills are all parameters that should be taken into account when designing the system, as they may affect users' perception of the system and attitude towards it along with their willingness to interact with it. According to one line of research, anthropomorphic, human-like robots promote universality, engagement, likeness, task efficiency [36, 53].

For practical systems, however, human-like appearance and behavior may trigger expectations that are not ultimately met. Unrestricted use of spoken language, mimicking emotions, humanoid appearance could elicit human-like responses that cannot be handled by current technology. Therefore, in this sense, it is important that the appearance and behaviour of the robot matches its abilities [38]. Furthermore, with regards to the robot's appearance, in particular, Butler and Agah (2001) showed that users favored smaller, "tecnomorphic" robots moving slowly, and approaching them indirectly; contrarily, large-size, humanoid robots were found to increase the level of user discomfort [40]. These results are corroborated by findings in another study according to which users disliked being directly, frontally approached by the mobile "fetch and carry" robot [41].

Also, according to Goetz et al. (2003), the successful design depends on the appropriate match between the robot's social skills/characteristics and its role in the task that it is designed for [42]. Based on their experiments, a machine-like approach was favored for more serious tasks such as security guards or lab assistants, while artistic and entertainment tasks called for a more human-like, playful, and emotional approach. According to this association, typical service robots in human-inhabited environments, such as mail delivery or floor cleaning robots, require little social skills, which could improve acceptance by the users.

Furthermore, user profiling is important for deciding upon the interaction strategy followed. Independently living elderly people, for example, maybe more interested in social interaction with a service robot given that they often live alone [1, 5] contrary to younger users who would place more significance on efficiency and task automatism. In short, social, life-like interfaces may not always be optimal interfaces, especially with regards to issues such as practical feasibility, effectiveness, and technology limitations.

Multimodality is another aspect that calls for attention [43]. For service robots, there is more to human-robot communication than verbal dialogue concerning the specification of tasks to be solved by the robot. The communication between humans and service robots can also be multimodal, incorporating verbal utterances, visual input and output, and perhaps gestures, position, and more.

All the above lead to specific approaches on the interaction (dialogue) management techniques that need to be employed, deployed, and tested in certain stages of the development.

6 Recent and Future Trends: Social Intelligence

As it has already been mentioned, the mere fact that service robots are now placed in dynamic, unstructured, and "socially oriented" environments operated by non-expert users calls for new models of interaction that build on collaborative and social skills.

According to Bartneck and Forlizzi's (2004) definition, "a social robot is an autonomous or semi-autonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact" [44]. More specifically, key behavior and appearance characteristics that indicate a robot's social intelligence are [36, 44, 45]:

- Display of personality traits such as politeness, seriousness, or playfulness.
- Compliance with social norms and rules specific to each society and culture. A robot receptionist is expected to exhibit behavior accordant to the established pattern for receptionists in the particular culture in terms of social distance, politeness, use of plural form, positioning, posture, and so forth.
- Interactivity, behavior adjustment according to the specific user and interaction setting, in response to external stimuli and contextual factors in general.
- Intelligent and intentional behavior, learning skills, decision making capability and autonomy, causal and collaborative behavior, awareness of human communication principles (e.g., turn-taking protocol, Grice's (1975) co-operative principle [46]).
- Employment of natural communication modalities such as spoken language and gesturing, facial expressions, eye contact, gazing, sensing touch.
- Posture adjustments, human-like movement (for example body part movement with varying velocity [36], appropriate positioning, talk and lip synchronization.
- Physical embodiment, based on the assumption that "life and intelligence only, develops inside a body" [47].
- Gender attribution, reference in the first person (e.g. "I'll get the coffee now" as opposed to "Getting the coffee…").
- Display and understanding of emotions, empathy.

In a study conducted by de Ruyter et al. (2005), social intelligence was shown to have a positive effect on user's perception and acceptance of the robot [36]. Human-like behavior was also shown to induce more social and collaborative behavior on the user's part. With regards to the latter, however, there is a certain degree of caution and reserve, as the underlying technology has not reached adequate maturity levels, and users may overestimate and over challenge robot's abilities, which would result in a decrease in efficiency and user satisfaction as user expectations are not met [38, 48]. In this line of thought, users are claimed to be more interested in practical characteristics as opposed to human-like characteristics. Nevertheless, interfaces that make use of at least some level of social skills and intelligence have been shown to be more enjoyable, trustworthy, usable, natural, engaging, and efficient [49–52].

Similarly, robotic systems exhibiting human-like emotive behavior as a particular aspect of social intelligence can increase user engagement and compliance, improve system acceptance, and facilitate decision making and learning processes, among others [21, 53]. In this respect, they are particularly appropriate as companions for the elderly, "game partners" or in areas such as e-learning and autism therapy. Emotions may be conveyed through facial expressions, eye-gaze and head movements, gesture and posture, touch, language/utterance content, appropriate prosody manipulation, and emotive vocalizations. Ultimately, robots should also be able to evaluate the emotional state of the user, indicated through any of the above modalities – as well as any other physiological signs such as heart rate [54] – and adjust their behavior accordingly [55, 56].

With regards to the speech modality, in particular, appropriate prosody manipulation is critical not only for affective, emotive interaction but also for displaying and communicating context-awareness. It is generally acknowledged that prosody is associated with the organization of information in an utterance, indicating how an utterance relates to the situational context [57]. For example, speakers may place pitch accents on different elements of the utterance in order to distinguish between new and given information (i.e., information that is already part of the common ground) or acknowledge the existence of alternative referents relevant to the entity under discussion (intonational contrast). Violation of intonation related grammar principles may lead to an ungrammatical, confusing, and unnatural spoken output. A model for the production of context-aware intonation in human-robot situated dialogue that is sensitive to such principles has been developed within the CogX project. The model assigns appropriate intonation patterns to convey properties such as contrast, theme-rheme distinction, uncertainty, and commitment [29] and, in this manner, support adaptation, and transparency in HRI.

7 Conclusion

This paper presented basic design principles and methodologies for the development of Spoken Dialogue Interfaces for service robots. Even more than traditional computer applications, the use of intelligent robots encourages the view of the machine as a partner in communication rather than as a tool. This suggests that people can be expected to apply more naturalness in the form of modalities and richness of interaction than in ordinary human-computer interfaces. As a result, SDIs that allow for voice as primary means of interaction have become central for the development of usable systems, especially taking into account that service robots are now typically operated by non-expert users to perform a variety of tasks in unstructured, dynamic environments.

Furthermore, contrary to on-screen, software agents or telephony-based spoken dialogue systems, mobile service robots must also reason about the spatial environment they operate in; the environment in which the robots act and the users live, the shared space between them, the location and the objects available shape the shared sub-world that the communication knowledge is drawn upon and complex use scenarios are formed. This situated form of human-robot interaction crucially affects standard design and usability evaluation methodologies, which must be adjusted to comply with these particular aspects of HRI. In this respect, design methodology should not be merely user-centered but situation and usability centered.

Similarly, usability evaluation approaches and metrics should be adjusted to appropriately address interface aspects important for HRI, such as physical embodiment, mobility, social relationships, collaboration, anthropomorphism, personalized communication, and multi-user interaction.

Acknowledgments. This work has been co-financed by the National and Kapodistrian University of Athens, Special Account for Research Grants.

References

1. Baltus, G., et al.: Towards personal service robots for the elderly. In: Workshop Interactive Robots and Entertainment (WIRE 2000) (2000). https://doi.org/10.1007/s12369-014-0232-4
2. Granata, C., Chetouani, M., Tapus, A., Bidaud, P., Dupourqué, V.: Voice and graphical-based interfaces for interaction with a robot dedicated to elderly and people with cognitive disorders. In: Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication (2010). https://doi.org/10.1109/ROMAN.2010.5598698
3. Fang, H., et al.: From captions to visual concepts and back. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2015). https://doi.org/10.1109/CVPR.2015.7298754
4. Kruijff, G.J.M., Zender, H., Jensfelt, P., Christensen, H.I.: Situated dialogue and understanding spatial organization: knowing what is where and what you can do there. In: Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication (2006). https://doi.org/10.1109/ROMAN.2006.314438
5. Severinson-Eklundh, K., Green, A., Hüttenrauch, H.: Social and collaborative aspects of interaction with a service robot. Robot. Auton. Syst. (2003). https://doi.org/10.1016/S0921-8890(02)00377-9
6. Stephanidis, C., Akoumianakis, D., Sfyrakis, M., Paramythis, A.: Universal accessibility in HCI : process-oriented design guidelines and tool requirements. In: 4th ERCIM Workshop User Interfaces All (1998)
7. Torta, E., Oberzaucher, J., Werner, F., Cuijpers, R.H., Juola, J.F.: Attitudes towards socially assistive robots in intelligent homes: results from laboratory studies and field trials. J. Hum.-Robot Interact. (2013). https://doi.org/10.5898/jhri.1.2.torta
8. Zobel, M., et al.: MOBSY: integration of vision and dialogue in service robots. In: Schiele, B., Sagerer, G. (eds.) ICVS 2001. LNCS, vol. 2095, pp. 50–62. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-48222-9_4
9. Tao, Y., Wei, H., Wang, T.: A speech interaction system based on finite state machine for service robot. In: Proceedings of the International Conference on Computer Science and Software Engineering, CSSE 2008 (2008). https://doi.org/10.1109/CSSE.2008.627
10. Matsui, T., et al.: Integrated natural spoken dialogue system of Jijo-2 mobile robot for office services. In: Proceedings of the National Conference on Artificial Intelligence (1999)
11. Wilske, S., Kruijff, G.J.: Service robots dealing with indirect speech acts. In: IEEE International Conference on Intelligent Robots and Systems (2006). https://doi.org/10.1109/IROS.2006.282259
12. Martin. D.J., Jurasky, D.: Speech and language processing: an introduction to natural language processing. In: SPEECH Language Processing An Introduction to Natural Language Processing Computational Linguistic Speech Recognition (2001)
13. Xydas, G., Spiliotopoulos, D., Kouroupetroglou, G.: Modeling prosodic structures in linguistically enriched environments. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 521–528. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30120-2_66
14. Spiliotopoulos, D., Xydas, G., Kouroupetroglou, G.: Diction based prosody modeling in table-to-speech synthesis. In: Matoušek, V., Mautner, P., Pavelka, T. (eds.) TSD 2005. LNCS (LNAI), vol. 3658, pp. 294–301. Springer, Heidelberg (2005). https://doi.org/10.1007/11551874_38

15. Spiliotopoulos, D., Androutsopoulos, I., Spyropoulos, C.D.: Human-robot interaction based on spoken natural language dialogue. In: Proceedings of the European Workshop on Service and Humanoid Robots, pp. 25–27 (2001)
16. Xydas, G., Spiliotopoulos, D., Kouroupetroglou, G.: Modeling emphatic events from non-speech aware documents in speech based user interfaces. In: Proceedings of Human Computer Interaction, pp. 806–810 (2003)
17. McTear, M.F.: Spoken dialogue technology: enabling the conversational user interface. ACM Comput. Surv. (2002). https://doi.org/10.1145/505282.505285
18. Burgard, W., et al.: Experiences with an interactive museum tour-guide robot. Artif. Intell. (1999). https://doi.org/10.1016/s0004-3702(99)00070-3
19. Siegwart, R., et al.: Robox at expo.02: a large-scale installation of personal robots. Robot. Auton. Syst. (2003). https://doi.org/10.1016/S0921-8890(02)00376-7
20. Dominey, P.F., Mallet, A., Yoshida, E.: Progress in programming the HRP-2 humanoid using spoken language. In: Proceedings of the IEEE International Conference on Robotics and Automation (2007). https://doi.org/10.1109/ROBOT.2007.363642
21. Picard, R.W.: Affective computing: challenges. Int. J. Hum Comput Stud. (2003). https://doi.org/10.1016/S1071-5819(03)00052-1
22. Hüttenrauch, H., Green, A., Norman, M., Oestreicher, L., Eklundh, K.S.: Involving users in the design of a mobile office robot. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. (2004). https://doi.org/10.1109/TSMCC.2004.826281
23. Tenbrink, T., Hui, S.: Negotiating spatial goals with a wheelchair. In: Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue (2007)
24. Kruijff, G.J.M., Zender, H., Jensfelt, P., Christensen, H.I.: Situated dialogue and spatial organization: what, where… and why? Int. J. Adv. Robot. Syst. (2007). https://doi.org/10.5772/5701
25. Leech, G.: Principles of Pragmatics (2016). https://doi.org/10.4324/9781315835976
26. Clark, H.H., Schaefer, E.F.: Contributing to discourse. Cogn. Sci. (1989). https://doi.org/10.1016/0364-0213(89)90008-6
27. Brennan, S.E., Hulteen, E.A.: Interaction and feedback in a spoken language system: a theoretical framework. Knowl.-Based Syst. (1995). https://doi.org/10.1016/0950-7051(95)98376-H
28. Steedman, M.: Information structure and the syntax-phonology interface. Linguist. Inq. (2000). https://doi.org/10.1162/002438900554505
29. Kruijff-Korbayová, I., Meena, R., Pyykkönen, P.: Perception of visual scene and intonation patterns of robot utterances. In: HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (2011). https://doi.org/10.1145/1957656.1957717
30. Cohen, M.H., Giangola, J.P., Balogh, J.: Introduction to voice user interfaces (2004)
31. Blandford, A.: Semi-structured qualitative studies. In: Encyclopedia Human-Computer Interaction (2013)
32. Yankelovich, N., Levow, G.A., Marx, M.: Designing speechActs: issues in speech user interfaces. In: Proceedings of the Conference on Human Factors in Computing Systems (1995)
33. Austin, J.L.: How to Do Things with Words. Harvard University Press, Cambridge (1975)
34. Allen, J.: Natural Language Understanding. Pearson, New Delhi (1995)
35. Fraser, N.M., Gilbert, G.N.: Simulating speech systems. Comput. Speech Lang. (1991). https://doi.org/10.1016/0885-2308(91)90019-M
36. De Ruyter, B., Saini, P., Markopoulos, P., Van Breemen, A.: Assessing the effects of building social intelligence in a robotic interface for the home. Interact. Comput. (2005). https://doi.org/10.1016/j.intcom.2005.03.003

37. Weinschenk, S., Barker, D.: Designing Effective Speech Interfaces. Wiley, Hoboken (2000)
38. Parlitz, C., Häagele, M., Klein, P., Seifert, J., Dautenhahn, K.: Care-O-bot 3 - rationale for human-robot interaction design. In: 39th International Symposium on Robotics, ISR 2008 (2008)
39. Takiguchi, T., Sako, A., Revaud, J., Yamagata, T., Miyake, N., Ariki, Y.: Human-robot interface using system request utterance detection based on acoustic features. In: Proceedings of the 2008 International Conference on Multimedia and Ubiquitous Engineering, MUE 2008 (2008). https://doi.org/10.1109/MUE.2008.87
40. Butler, J.T., Agah, A.: Psychological effects of behavior patterns of a mobile personal robot. Auton. Robots (2001). https://doi.org/10.1023/A:1008986004181
41. Dautenhahn, K., et al.: How may I serve you? A robot companion approaching a seated person in a helping context. In: HRI 2006: Proceedings of the 2006 ACM Conference on Human-Robot Interaction (2006)
42. Goetz, J., Kiesler, S., Powers, A.: Matching robot appearance and behavior to tasks to improve human-robot cooperation. In: Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication (2003). https://doi.org/10.1109/ROMAN.2003.1251796
43. Alexandersson, J., et al.: Metalogue: A multiperspective multimodal dialogue system with metacognitive abilities for highly adaptive and flexible dialogue management. In: Proceedings of the 2014 International Conference on Intelligent Environments, IE 2014, pp. 365–368 (2014). https://doi.org/10.1109/IE.2014.67
44. Bartneck, C., Forlizzi, J.: A design-centred framework for social human-robot interaction. In: Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication (2004). https://doi.org/10.1109/roman.2004.1374827
45. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robot. Auton. Syst. (2003). https://doi.org/10.1016/S0921-8890(02)00372-X
46. Grice, H.P.: Logic and conversation. In: Syntax and Semantics. Speech Arts, vol. 3 (1975)
47. Dautenhahn, K.: Embodiment and interaction in socially intelligent life-like agents. In: Nehaniv, C.L. (ed.) CMAA 1998. LNCS (LNAI), vol. 1562, pp. 102–141. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-48834-0_7
48. Pearson, J., Hu, J., Branigan, H.P., Pickering, M.J., Nass, C.I.: Adaptive language behavior in HCl: how expectations and beliefs about a system affect users' word choice. In: Proceedings of the Conference on Human Factors in Computing Systems (2006)
49. Bickmore, T., Cassell, J.: Relational agents: a model and implementation of building user trust. In: Proceedings of the Conference on Human Factors in Computing Systems (2001)
50. Heylen, D., Es, I., Nijholt, A., Dijk, E.: Experimenting with the gaze of a conversational agent. In: Proceedings of the International CLASS Workshop Natural Intelligent and Effective Interaction Multimodal Dialogue Systems (2002)
51. Bartneck, C.: Interacting with an embodied emotional character. In: Proceedings of the International Conference on Designing Pleasurable Products and Interfaces (2003). https://doi.org/10.1145/782910.782911
52. Bruce, A., Nourbakhsh, I., Simmons, R.: The role of expressiveness and attention in human-robot interaction. In: Proceedings of the IEEE International Conference on Robotics and Automation (2002). https://doi.org/10.1109/robot.2002.1014396
53. Breazeal, C.: Affective interaction between humans and robots. In: Kelemen, J., Sosík, P. (eds.) ECAL 2001. LNCS (LNAI), vol. 2159, pp. 582–591. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44811-X_66
54. Kulíc, D., Croft, E.: Affective state estimation for human-robot interaction. In: IEEE Trans. Robot. (2007). https://doi.org/10.1109/TRO.2007.904899

55. Breazeal, C.: Designing sociable robots. In: Designing Sociable Robots (2018). https://doi.org/10.7551/mitpress/2376.003.0007
56. Dautenhahn, K.: Socially intelligent agents - the human in the loop (2001). https://doi.org/10.1109/TSMCA.2001.952709
57. Spiliotopoulos, D., Xydas, G., Kouroupetroglou, G., Argyropoulos, V., Ikospentaki, K.: Auditory universal accessibility of data tables using naturally derived prosody specification. Univ. Access Inf. Soc. 9 (2010). https://doi.org/10.1007/s10209-009-0165-0