A review of verbal and non-verbal human–robot interactive communication

CrossMark

Nikolaos Mavridis *

Interactive Robots and Media Lab, NCSR Demokritos, GR-15310, Agia Paraskevi, Athens, Greece

HIGHLIGHTS

- An extensive survey of human–robot interactive communication is provided.
- Verbal as well as non-verbal aspects of human–robot interaction are covered.
- The paper starts with a historical introduction.
- Ten special desiderata that human–robot systems should fulfil are proposed.
- The desiderata are examined in detail, culminating to a forward-looking discussion.

ARTICLE INFO

ABSTRACT

In this paper, an overview of human–robot interactive communication is presented, covering verbal as well as non-verbal aspects. Following a historical introduction, and motivation towards fluid human–robot communication, ten desiderata are proposed, which provide an organizational axis both of recent as well as of future research on human–robot communication. Then, the ten desiderata are examined in detail, culminating in a unifying discussion, and a forward-looking conclusion.

1. Introduction: historical overview

While the first modern-day industrial robot, Unimate, began work on the General Motors assembly line in 1961, and was conceived in 1954 by George Devol [1,2], the concept of a robot has a very long history, starting in mythology and folklore, and the first mechanical predecessors (automata) having been constructed in Ancient Times. For example, in Greek mythology, the God Hephaestus is reputed to have made mechanical servants from gold ([3] in p. 114, and [4] verse 18.419). Furthermore, a rich tradition of designing and building mechanical, pneumatic or hydraulic automata also exists: from the automata of Ancient Egyptian temples, to the mechanical pigeon of the Pythagorean Archytas of Tarantum circa 400 BC [5], to the accounts of earlier automata found in the Lie Zi text in China in 300 BC [6], to the devices of Heron of Alexandria [7] in the 1st century. The Islamic world also plays an important role in the development of automata; Al-Jazari, an Arab inventor, designed and constructed numerous automatic machines, and is even reputed to have devised the first programmable humanoid robot in 1206 AD [8]. The word "robot", a Slavic word meaning servitude, was first used in this context by the Czech author Karel Capek in 1921 [9].

However, regarding robots with natural-language conversational abilities, it was not until the 1990s that the first pioneering systems started to appear. Despite the long history of mythology and automata, and the fact that even the mythological handmaidens of Hephaestus were reputed to have been given a voice [3], and despite the fact that the first general-purpose electronic speech synthesizer was developed by Noriko Omeda in Japan in 1968 [10], it was not until the early 1990s that conversational robots such as MAIA [11], RHINO [12], and AESOP [13] appeared. These robots cover a range of intended application domains; for example, MAIA was intended to carry objects and deliver them, while RHINO is a museum guide robot, and AESOP a surgical robot.

In more detail, the early systems include Polly, a robotic guide that could give tours in offices [14,15]. Polly had very simple interaction capacities; it could perceive human feet waving a "tour

---

* Tel.: +30 2310824077.
E-mail address: nmav@alum.mit.edu.

wanted" signal, and then it would just use pre-determined phrases during the tour itself. A slightly more advanced system was TJ [16]. TJ could verbally respond to simple commands, such as "go left", albeit through a keyboard. RHINO, on the other hand [12], could respond to tour-start commands, but then, again, just offered a pre-programmed tour with fixed programmer-defined verbal descriptions. Regarding mobile assistant robots with conversational capabilities in the 1990s, a classic system is MAIA [11,17], obeying simple commands, and carrying objects around places, as well as the mobile office assistant which could not only deliver parcels but also guide visitors described in [18], and the similar in functionality Japanese-language robot Jijo-2 [19–21]. Finally, an important book from the period is [22], which is characteristic of the traditional natural-language semantics-inspired theoretical approaches to the problem of human–robot communication, and also of the great gap between the theoretical proposals and the actual implemented systems of this early decade.

What is common to all the above early systems is that they share a number of limitations. First, all of them only accept a fixed and small number of simple *canned commands*, and they respond with a set of *canned answers*. Second, the only *speech acts* (in the sense of Searle [23]) that they can handle are requests. Third, the dialogue they support is clearly not flexibly *mixed initiative*; in most cases it is just human-initiative. Four, they do not really support *situated language*, i.e. language about their physical situations and events that are happening around them; except for a fixed number of canned location names in a few cases. Five, they are not able to handle *affective speech*; i.e. emotion-carrying prosody is neither recognized nor generated. Six, their *non-verbal communication* [24] capabilities are almost non-existent; for example, gestures, gait, facial expressions, and head nods are neither recognized nor produced. And seventh, their dialogue systems are usually effectively stimulus–response or stimulus-state-response systems; i.e. no real *speech planning* or purposeful dialogue generation is taking place, and certainly not in conjunction with the motor planning subsystems of the robot. Last but quite importantly, no real *learning*, off-line or on-the-fly is taking place in these systems; verbal behaviours have to be prescribed.

All of these shortcomings of the early systems of the 1990s, effectively have become desiderata for the next two decades of research: the 2000s and 2010s, which we are in at the moment. Thus, in this paper, we will start by providing a discussion giving motivation to the need for existence of interactive robots with natural human–robot communication capabilities, and then we will enlist a number of desiderata for such systems, which have also effectively become areas of active research in the last decade. Then, we will examine these desiderata one by one, and discuss the research that has taken place towards their fulfilment. Special consideration will be given to the so-called "symbol grounding problem" [25], which is central to most endeavours towards natural language communication with physically embodied agents, such as robots. Finally, after a discussion of the most important open problems for the future, we will provide a concise conclusion.

2. Motivation: interactive robots with natural language capabilities but why?

There are at least two avenues towards answering this fundamental question, and both will be attempted here. The first avenue will attempt to start from first principles and derive a rationale towards equipping robots with natural language. The second, more traditional and safe avenue, will start from a concrete, yet partially transient, base: application domains existing or potential. In more detail:

Traditionally, there used to be a clear separation between design and deployment phases for robots. Application-specific robots (for example, manufacturing robots, such as [26]) were: (a) designed by expert designers, (b) possibly tailor-programmed and occasionally reprogrammed by specialist engineers at their installation site, and (c) interacted with their environment as well as with specialized operators during actual operation. However, not only the phenomenal simplicity but also the accompanying inflexibility and cost of this traditional setting is often changing nowadays. For example, one might want to have broader-domain and less application-specific robots, necessitating more generic designs, as well as less effort by the programmer-engineers on site, in order to cover the various contexts of operation. Even better, one might want to rely less on specialized operators, and to have robots interact and collaborate with non-expert humans with a little if any prior training. Ideally, even the actual traditional programming and re-programming might also be transferred over to non-expert humans; and instead of programming in a technical language, to be replaced by intuitive tuition by demonstration, imitation and explanation [27–29]. Learning by demonstration and imitation for robots already has quite some active research; but most examples only cover motor and aspects of learning, and language and communication is not involved deeply.

And this is exactly where natural language and other forms of fluid and natural human–robot communication enter the picture: Unspecialized non-expert humans are used to (and quite good at) teaching and interacting with other humans through a mixture of natural language as well as nonverbal signs. Thus, it makes sense to capitalize on this existing ability of non-expert humans by building robots that do not require humans to adapt to them in a special way, and which can fluidly collaborate with other humans, interacting with them and being taught by them in a natural manner, almost as if they were other humans themselves.

Thus, based on the above observations, the following is one classic line of motivation towards justifying efforts for equipping robots with natural language capabilities: why not build robots that can comprehend and generate human-like interactive behaviours, so that they can cooperate with and be taught by non-expert humans, so that they can be applied in a wide range of contexts with ease? And of course, as natural language plays a very important role within these behaviours, why not build robots that can fluidly converse with humans in natural language, also supporting crucial non-verbal communication aspects, in order to maximize communication effectiveness, and enable their quick and effective application?

Thus, having presented the classical line of reasoning arriving towards the utility of equipping robots with natural language capabilities, and having discussed a space of possibilities regarding role assignment between human and robot, let us now move to the second, more concrete, albeit less general avenue towards justifying conversational robots: namely, specific applications, existing or potential. Such applications, where natural human–robot interaction capabilities with verbal and non-verbal aspects would be desirable, include: flexible manufacturing robots; lab or household robotic assistants [30–33]; assistive robotics and companions for special groups of people [34]; persuasive robotics (for example, [35,36]); robotic receptionists [37], robotic educational assistants, shopping mall robots [38], museum robots [39,40], tour guides [41,42], environmental monitoring robots [43], robotic wheelchairs [44,45], companion robots [46], social drink-serving robots [47], all the way to more exotic domains, such as robotic theatre actors [48,49], musicians [50], and dancers [51].

In almost all of the above applications, although there is quite some variation regarding requirements, one aspect at least is shared: the desirability of natural fluid interaction with humans supporting natural language and non-verbal communication, possibly augmented with other means. Of course, although this might be desired, it is not always justified as the optimum choice, given

techno-economic constraints of every specific application setting. A thorough analysis of such constraints together with a set of guidelines for deciding when natural-language interaction is justified, can be found in [52].

Now, having examined justifications towards the need for natural language and other human-like communication capabilities in robots across two avenues, let us proceed and become more specific: natural language, indeed but what capabilities do we actually need?

3. Desiderata—what might one need from a conversational robot?

An initial list of desiderata is presented below, which is neither totally exhaustive nor absolutely orthogonal; however, it serves as a good starting point for discussing the state of the art, as well as the potentials of each of the items:

(D1) Breaking the "simple commands only" barrier.
(D2) Multiple speech acts.
(D3) Mixed initiative dialogue.
(D4) Situated language and the symbol grounding problem.
(D5) Affective interaction.
(D6) Motor correlates and Non-Verbal Communication.
(D7) Purposeful speech and planning.
(D8) Multi-level learning.
(D9) Utilization of online resources and services.
(D10) Miscellaneous abilities.

The particular order of the sequence of desiderata, was chosen for the purpose of illustration, as it provides partially for a building-up of key points, also allowing for some tangential deviations. Not all desiderata are necessarily of equal difficulty, and arguably D1, D3–4, and D7–8 have so far proven to be particularly hard. One of the main reasons underlying this situation has to do with the divide between the two worlds that interactive robots usually live in: the symbolic/discrete world of logical representations and language on the one hand, and the continuous and noisy world of sensorymotor data on the other. And it is not only the uncertainty that arises from the unreliability of the sensorymotor end that contributes to the difficulties, but also the fact that sensor data tends to be structured in ways that are not easily alignable to the requirements of symbolic representations, as we shall see. Let us now proceed and examine the desiderata in detail one by one:

3.1. Breaking the "simple commands only" barrier

The traditional conception of conversational robots, as well as most early systems, is based on a clear human-master robot-servant role assignment, and restricts the robots conversational competencies to simple "motor command requests" only in most cases. For example, in systems such as [30,53], a typical dialogue might be:

H: "Give me the red one".
R: (Picks up the red ball, and gives to human.)
H: "Give me the green one".
R: "Do you mean this one, or that one?" (robot points to two possible candidate objects).
H: "The one on the left".
R: (Picks up the green ball on the left, and hands over to human.)

What are the main points noticed in this example? Well, first of all, (p1) this is primarily a single-initiative dialogue: the human drives the conversation, the robot effectively just produces motor and verbal responses to the human verbal stimulus. Second, (p2) apart from some disambiguating questions accompanied

by deixis, there is not much that the robot says the robot primarily responds with motor actions to the human requests, and does not speak. And, (p3) regarding the human statements, we only have one type of speech acts [23]: RequestForMotorAction. Furthermore, (p4) usually such systems are quite inflexible regarding multiple surface realizations of the acceptable commands; i.e. the human is allowed to say "Give me the red one", but if he instead used the elliptical "the red object, please" he might have been misinterpreted and (p5) in most cases, the mapping of words-to-responses is arbitrarily chosen by the designer; i.e. motor verbs translate to what the designer thinks they should mean for the robot (normative meaning), instead of what an empirical investigation would show regarding what other humans would expect they mean (empirical meaning).

Historically, advanced theorization for such systems exists as early as [22]. Actually, if one extends from physical robots to systems comprising a virtual robot in a virtual world, Winograds SHRDLU program [54,55] from the early seventies could already support multiple speech acts and basic mixed initiative dialogue. There is still quite a stream of active research which, although based on beautiful and systematic formalizations and eloquent grammars, basically produces systems which would still fall within the three points mentioned above. Such an example is [56], in which a mobile robot in a multi-room environment, can handle commands such as: "Go to the breakroom and report the location of the blue box".

Notice that here we are not claiming that there is no importance in this research that falls within this strand; we are just mentioning that, as we shall see, there are many other aspects of natural language and robots, which are left unaccounted by such systems. Furthermore, it remains to be seen, how many of these aspects can later be effectively integrated with systems belonging to this strand of research.

3.2. Multiple speech acts

The limitations (p1)–(p5) cited above for the classic "simple commands only" systems provide useful departure points for extensions. Speech act theory was introduced by J.L. Austin [57], and a speech act is usually defined as an utterance that has performative function in language and communication. Thus, we are focusing on the function and purpose of the utterance, instead of the content and form. Several taxonomies of utterances can be derived according to such a viewpoint: for example, Searle [58], proposed a classification of illocutionary speech acts into assertives, directives, commisives, expressives, and declarations. Computational models of speech acts have been proposed for use in human–computer interaction [59].

In the light of speech acts, lets us start by extending upon point (p3) made in the previous section. In the short human–robot dialogue presented in the previous section, the human utterances "Give me the red one" and "Give me the green one" could be classified as Request speech acts, and more specifically requests for motor action (one could also have requests for information, such as "What colour is the object?"). But what else might one desire in terms of speech act handling capabilities, apart from RequestForMotorAction (which we shall call SA1, a Directive according to [58])? Some possibilities follow below:

H: "How big is the green one?" (RequestForInformAct, SA2, Directive).
H: "There is a red object at the left" (Inform, SA3, Assertive).
H: "Let us call the small doll Daisy" (Declare, SA4, Declaration).

And many more exist. Systems such as [53] are able to handle SA2 and SA3 apart from SA1-type acts; and one should also notice, that there are many classificatory systems for speech acts,

across different axes of classification, and with multiple granularities. Also, it is worth starting at this stage to contemplate upon what might it mean to respond appropriately to different kinds of speech acts. For example, an appropriate response to a RequestForMotorAction (a Directive) is the motor action itself, if unambiguous and feasible; however, an appropriate response to an Assertive or a Declarative consists of a change to some form of a "mental model" [60] or "situation model" [61,53] that the robot might be keeping; i.e. creating an appropriate mental token for an object in the case of "There is a red object at the left", or changing the name label for a mental object token in the case of "Let us call this small doll Daisy"; i.e. both statements elicit primarily internal (mental) actions, instead of external (motor or verbal) actions.

Another relevant aspect of speech act theory is the handling of indirect speech acts. For example, consider the following utterance:

H: "Ah, it is quite hot in this room" (phenomenally, an Assertive),
    which might actually be a polite way of saying:
H: "Open the window" (essentially, a Directive).

This substitution of an Assertive for an implied Directive (to be inferred by the listener) is a classic example of an indirect speech act. Usually, the analysis of such acts is based on the Gricean maxims of conversation [62]; and numerous computational implementations for handling such indirect speech acts have been proposed, such as [63].

Finally, yet another problem related to speech acts, is the issue of their classification from the robot, after hearing them. Classic techniques such as those described in [64] rely on linguistic information only; however, paralinguistic information (such as prosodic features) can also prove useful towards speech act classification; the interested reader is referred for example to [65].

3.3. Mixed initiative dialogue

Now, starting again from the shortcoming of the traditional "simple commands-only" systems, let us extend across another axis, namely (p1): human-initiative dialogue is not the only option; one could also have robot-initiative, or ideally, full mixed-initiative. Consider FaceBots [66,67], a conversational robot utilizing facebook-derived information. A typical dialogue might include turns such as:

R: "Hello! Are you Noura Dhaheri?"
H: "Yes".
R: "Hello Noura! Are you doing well today?"
H: "Pretty well!"
R: "Great! Do you know that our friend Wajahat Kazmi has posted a new photo?"
H: "No!"
R: "Well, I am sending you an inbox message for this, check it out!"

This is a classic example of robot-initiative dialogue, and actually with very simple responses expected from the human mainly Yes or No. On the other hand, consider a dialogue such as (from the robot BIRON at Bielefeld, dialogue in [68]), which is an example of a dialogue with limited mixed initiative:

R: "Hello Human!"
H: "Hello! Who are you?"
R: "My name is BIRON, if you show me something I can remember it".
H: "OK, um, look, this is a mug".
R: "I beg your pardon?"
H: "This is a mug, umm, here".
R: "Sorry, can you use some other words?"
H: "This is a mug!"
R: "Sorry, I cannot find it. Are you pointing to it with your finger?"

H: "Yes!"
R: "I know it is sometimes difficult with me, but dont feel discouraged!"
H: (laugh).
R: "OK, I have found it, it is really nice!"

Here, it is neither the robot nor the human driving the dialogue all of the time; for example, the opening pair, R–H is robot-initiative (R: "Hello Human!" H: "Hello!"). However, directly afterwards, the second half of H together with R creates a second pair of exchanges, which is human-initiative (H: "Who are you?", R: "My names is BIRON…"). And thus the initiative can be reversed in multiple points throughout the dialogue.

For an investigation of the state of the art towards mixed initiative, the interested reader is referred to examples such as the Karlsruhe Humanoid [69] the Biron and Barthoc systems at Bielefeld [68], and also workshops such as [70].

3.4. Situated language and symbol grounding

Yet another observation regarding shortcomings of the traditional command-only systems that are worth extending from, was point (p5) that was mentioned above: the meanings of the utterances were normatively decided by the designer, and not based on empirical observations. For example, a designer/coder could normatively pre-define the semantics of the colour descriptor "red" as belonging to the range between two specific given values. Alternatively, one could empirically get a model of the applicability of the descriptor "red" based on actual human usage; by observing the human usage of the word in conjunction with the actual apparent colour wavelength and the context of the situation. Furthermore, the actual vocabularies (red, "pink", etc.) or the classes of multiple surface realizations (p4) (quasi-synonyms or semantically equivalent parts of utterances, for example: "give me the red object", "hand me the red ball"), are usually hand-crafted in such systems, and again not based on systematic human observation or experiment.

There are a number of notable exceptions to this rule, and there is a growing tendency to indeed overcome these two limitations recently. For example, consider [71], during which a wizard-of-oz experiment provided the collection of vocabulary from users desiring to verbally interact with a robotic arm, and examples such as [44], for which the actual context-depending action models corresponding to simple verbal commands like "go left" or "go right" (which might have quite different expected actions, depending on the surrounding environment) were learnt empirically through human experiments.

Embarking upon this avenue of thought, it slowly becomes apparent that the connection between local environment (and more generally, situational context) and procedural semantics of an utterance is quite crucial. Thus, when dealing with robots and language, it is impossible to isolate the linguistic subsystems from perception and action, and just plug-and-play with a simple speech-in speech-out black box chatterbot of some sort (such as the celebrated ELIZA [72] or even the more recent victors of the Loebner Prize [73]). Simply put, in such systems, there is no connection of what is being heard or said to what the robot senses and what the robot does. This is quite a crucial point; there is a fundamental need for closer integration of language with sensing, action, and purpose in conversational robots [30,53], as we shall also see in the next sections.

3.4.1. Situated language

Upon discussing the connection of language to the physical context, another important concept becomes relevant: situated language, and especially the language that children primarily use

during their early years; i.e. language that is not abstract or about past or imagined events; but rather concrete, and about the physical here-and-now. But what is the relevance of this observation to conversational robots? One possibility is the following; given that there seems to be a progression of increasing complexity regarding human linguistic development, often in parallel to a progression of cognitive abilities, it seems reasonable to: first partially mimic the human developmental pathway, and thus start by building robots that can handle such situated language, before moving on to a wider spectrum of linguistic abilities. This is for example the approach taken at [53].

Choosing situated language as a starting point also creates a suitable entry point for discussing language grounding in the next section. Now, another question that naturally follows is: could one postulate a number of levels of extensions from language about the concrete here-and-now to wider domains? This is attempted in [53], and the levels of increasing detachment from the "here-and-now" postulated there are:

First level: limited only to the "here-and-now, existing concrete things". Words connect to things directly accessible to the senses at the present moment. If there is a chair behind me, although I might have seen it before, I cannot talk about it—"out of sight" means "non-existing" in this case. For example, such a robotic system is [74].

Second level: ("now, existing concrete things"); we can talk about the "now", but we are not necessarily limited to the "here"—where here means currently accessible to the senses. We can talk about things that have come to our senses previously, that we conjecture still exist through some form of psychological "object permanence" [75]—i.e., we are keeping some primitive "mental map" of the environment. For example, this was the state of the robot Ripley during [76].

Third level: ("past or present, existing concrete things"), we are also dropping the requirement of the "now"—in this case, we also possess some form of episodic memory [77] enabling us to talk about past states. An example robot implementation can be found in [78].

Fourth level: ("imagined or predicted concrete things"); we are dropping the requirement of actual past or present existence, and we can talk about things with the possibility of actual existence—either predicted (connectible to the present) or imagined [53].

Fifth level: ("abstract things") we are not talking about potentially existing concrete things any more, but about entities that are abstract. But what is the criterion of "concreteness?" A rough possibility is the following: a concrete thing is a first-order entity (one that is directly connected to the senses); an "abstract" thing is built upon first order entities, and does not connect directly to the senses, as it deals with relationships between them. Take, for example, the concept of the "number three": it can be found in an auditory example ("threeness" in the sound of three consecutive ticks); it can also be found in a visual example ("threeness" in the snapshot of three birds sitting on a wire). Thus, threeness seems to be an abstract thing (not directly connected to the senses).

Currently, there exist robots and methodologies [53] that can create systems handling basic language corresponding to the first four stages of detachment from situatedness; however, the fifth seems to still be out of reach. If what we are aiming towards is a robot with a deeper understanding of the meaning of words referring to abstract concepts, although related work on computational analogy making (such as [79]), could prove to provide some starting points for extensions towards such domains, we are still beyond the current state-of-the-art.

Nevertheless, there are two interesting points that have arisen in the previous sections: first, that when discussing natural language and robots, there is a need to connect language not only to sensory data, but also to internalized "mental models" of the world

in order for example to deal with detachment from the immediate "here-and-now". And second, that one needs to consider not only phonological and syntactical levels of language but also questions of semantics and meaning; and pose the question: "what does it mean for a robot to understand a word that it hears or utters?" And also, more practically: what are viable computational models of the meaning of words, suitable to embodied conversational robots? We will try to tackle these questions right now, in the next subsection.

3.4.2. Symbol grounding

One of the main philosophical problems that arises when trying to create embodied conversational robots is the so-called "symbol grounding problem" [25]. In simple terms, the problem is the following: imagine a robot, having an apple in front of it, and hearing the word "apple" a verbal label which is a conventional sign (in semiotic terms [80,81]), and which is represented by a symbol within the robots cognitive system. Now this sign is not irrelevant to the actual physical situation; the human that uttered the word "apple" was using it to refer to the physical apple that is in front of the robot. Now the problem that arises is the following: how can we connect the symbol standing for "apple" in the robots cognitive system, with the physical apple that it refers to? Or, in other words, how can we ground out the meaning of the symbol to the world? In simple terms, this is an example of the symbol grounding problem. Of course, it extends not only to objects signified by nouns, but to properties, relations, events, etc., and there are many other extensions and variations of it.

So, what are solutions relevant to the problem? In the case of embodied robots, the connection between the internal cognitive system of the robot (where the sign is) and the external world (where the referent is) is mediated through the sensory system, for this simple case described above. Thus, in order to ground out the meaning, one needs to connect the symbol to the sensory data say, to vision. Which is at least, to find a mechanism through which, achieves the following bidirectional connection: first, when an apple appears in the visual stream, instantiates an apple symbol in the cognitive system (which can later for example trigger the production of the word "apple" by the robot), and second, when an apple symbol is instantiated in the cognitive system (for example, because the robot heard that "there is an apple"), creates an expectation regarding the contents of the sensory stream given that an apple is reported to be present. This bidirectional connection can be succinctly summarized as:

external referent > sensory stream > internal symbol > produced utterance

external referent < sensory expectation < internal symbol < heard utterance.

This bidirectional connection we will refer to as "full grounding", while its first unidirectional part as "half grounding". Some notable papers presenting computational solutions of the symbol grounding problem for the case of robots are: half-grounding of colour and shapes for the Toco robot [74], and full-grounding of multiple properties for the Ripley robot [30]. Highly relevant work includes: [82] and also Steels [83–85], and also [86] from a child lexical perspective.

The case of grounding of spatial relations (such as "to the left of", and "inside") reserves special attention, as it is a significant field on its own. A classic paper is [87], presenting an empirical study modelling the effect of central and proximal distance on 2D spatial relations; regarding the generation and interpretation of referring expressions on the basis of landmarks for a simple rectangle world, there is [88], while the book by [89] extends well into illustrating the inadequacy of geometrical models and the need for functional models when grounding terms such as "inside", and covers a range of relevant interesting subjects. Furthermore, regarding the grounding of attachment and support relations in videos, there is

the classic work by [90]. For an overview of recent spatial semantics research, the interested reader is referred to [91], and a sampler of important current work in robotics includes [92–94], and the most recent work of Tellex on grounding with probabilistic graphical models [95], and for learning word meanings from unaligned parallel data [96].

Finally, an interesting question arises when trying to ground out personal pronouns, such as "me, my, you, your". Regarding their use as modifiers of spatial terms ("my left"), relevant work on a real robot is [76], and regarding more general models of their meaning, the reader is referred to [97], where a system learns the semantics of the pronouns through examples.

A number of papers have recently also appeared claiming to have provided a solution to the "symbol grounding problem", such as [98]. There is a variety of different opinions regarding what an adequate solution should accomplish, though. A stream of work around an approach dealing with the evolution of language and semiotics, is outlined in [99]. From a more applied and practical point of view though, one would like to be able to have grounded ontologies [100,101] or even robot-useable lexica augmented with computational models providing such grounding: and this is the ultimate goal of the EU projects POETICON [102,103], and the follow-up project POETICON II.

Another important aspect regarding grounding is the set of qualitatively different possible target meaning spaces for a concept. For example, [53] proposes three different types of meaning spaces: sensory, sensorimotor, and teleological. A number of other proposals exist for meaning spaces in cognitive science, but not directly related to grounding; for example, the geometrical spaces proposal of Gardenfors [104]. Furthermore, any long-ranging agenda towards extending symbol grounding to an ever-increasing range of concepts, needs to address yet another important point: semantic composition, i.e. for a very simple example, consider how a robot could combine a model of "red" with a model of "dark" in order to derive a model of "dark red". Although this is a fundamental issue, as discussed in [53], it has yet to be addressed properly.

Last but not least, regarding the real-world acquisition of large-scale models of grounding in practice, special data-driven models are required, and the quantities of empirical data required would make collection of such data from non-experts (ideally online) highly desirable. Towards that direction, there exists the pioneering work of Gorniak [85] where a specially modified computer game allowed the collection of referential and functional models of meaning of the utterances used by the human players. This was followed up by [105–107], in which specially designed online games allowed the acquisition of scripts for situationally appropriate dialogue production. These experiments can be seen as a special form of crowdsourcing, building upon the ideas started by pioneering systems such as Luis Von Ahns peekaboom game [108], but especially targeting the situated dialogic capabilities of embodied agents. Much more remains to be done in this promising direction in the future.

3.4.3. Meaning negotiation

Having introduced the concept of non-logic-like grounded models of meaning, another interesting complication arises. Given that different conversational partners might have different models of meaning, say for the lexical semantics of a colour term such as "pink", how is communication possible? A short, yet minimally informative answer, would be: given enough overlap of the particular models, there should be enough shared meaning for communication. But if one examines a number of typical cases of misalignment across models, he will soon reach to the realization that models of meaning, or even second-level models (beliefs about the models that others hold), are very often being negotiated and adjusted online, during a conversation. For example:

(Turquoise object on robot table, in front of human and robot)

H: "Give me the blue object!"
R: "No such object exists".
H: "Give me the blue one!"
R: "No such object exists".

But why is this surreal human–robot dialogue taking place, and why it would not have taken place for the case of two humans in a similar setting? Let us analyse the situation. The object on the table is turquoise, a colour which some people might classify as "blue", and others as "green". The robots colour classifier has learnt to treat turquoise as green; the human classifies the object as "blue". Thus, we have a categorical misalignment error, as defined in [53]. For the case of two humans interacting instead of a human and a robot, given the non-existence of another unique referent satisfying the "blue object" description, the second human would have readily assumed that most probably the first human is classifying turquoise as "blue"; and, thus, he would have temporarily adjusted his model of meaning for "blue" in order to be able to include turquoise as "blue", and thus to align his communication with his conversational partner. Thus, ideally we would like to have conversational robots that can gracefully recover from such situations, and fluidly negotiate their models of meaning online, in order to be able to account for such situations. Once again, this is a yet unexplored, yet crucial and highly promising avenue for future research.

3.5. Affective interaction

An important dimension of cognition is the affective/emotional. In the german psychological tradition of the 18th century, the affective was part of the tripartite classification of mental activities into cognition, affection, and conation; and apart from the widespread use of the term, the influence of the tri-partite division extended well into the 20th century [109].

The affective dimension is very important in human interaction [110], because it is strongly intertwined with learning [111], persuasion [112], and empathy, among many other functions. Thus, it carries over its high significance for the case of human–robot interaction. For the case of speech, affect is marked both in the semantic/pragmatic content as well as in the prosody of speech: and thus both of these ideally need to be covered for effective human–robot interaction, and also from both the generation as well as recognition perspectives. Furthermore, other affective markers include facial expressions, body posture and gait, as well as markers more directly linked to physiology, such as heart rate, breathing rate, and galvanic skin response.

Pioneering work towards affective human–robot interaction includes [113] where, extending upon analogous research from virtual avatars such as Rea [114], Steve [115], and Greta [116], Cynthia Breazeal presents an interactive emotion and drive system for the Kismet robot [117], which is capable of multiple facial expressions. An interesting cross-linguistic emotional speech corpus arising from children's interactions with the Sony AIBO robot is presented in [118]. Another example of preliminary work based on a Wizard-of-Oz approach, this time regarding children's interactions with the ATR Robovie robot in Japan, is presented in [119]. In this paper, automatic recognition of embarrassment or pleasure of the children is demonstrated. Regarding interactive affective storytelling with robots with generation and recognition of facial expressions, [120] present a promising starting point. Recognition of human facial expressions is accomplished through SHORE [121], as well as the Seeing Machines product FaceAPI. Other available facial expression recognition systems include [122], which has also been used as an aid for autistic children, as well as [123,124], where the output of the system is at the level of facial action coding (FACS).

Regarding generation of facial expressions for robots, some examples of current research include [125–127]. Apart from static poses, the dynamics of facial expressions is also very important towards conveying believability; for empirical research on dynamics see for example [128]. Still, compared to the wealth of available research on the same subject with virtual avatars, there is still a lag both in empirical evaluations of human–robot affective interaction, as well as in importing existing tools from avatar animation towards their use for robots.

Regarding some basic supporting technologies of affect-enabled text-to-speech and speech recognition, the interested reader can refer to the general reviews by Schroeder [129] on TTS, and by Ververidis and Kotropoulos [130] on recognition. A wealth of other papers on the subject exist; with some notable developments for affective speech-enabled real-world robotic systems including [131,132]. Furthermore, if one moves beyond prosodic affect, to semantic content, the wide literature on sentiment analysis and shallow identification of affect applies directly; for example [133–135]. Regarding physiological measurables, products such as Affectivas Q sensor [136], or techniques for measuring heart rate, breathing rate, galvanic skin response and more, could well become applicable to the human–robot affective interaction domain, of course under the caveats of [137]. Yet another important question for which still many aspects remain unanswered, is concerned with conveying emotions for the case of robotic embodiments which are not anthropomorphic and do not support speech, an initial investigation of which is presented in [138,139], and an example empirical study for the case of the emotions conveyed by a UAV can be found in [140]. Also, another interesting option is concerned with utilizing non-linguistic utterances (NLU) for conveying emotion, i.e. non-verbal sounds, as is done in [141], which are interpreted categorically [142]. Finally, it is worth noting that significant cross-culture variation exists regarding affect; both at the generation, as well as at the understanding and situational appropriateness levels [143]. In general, affective human–robot interaction is a growing field with promising results, which is expected to grow even more in the near future.

3.6. Motor correlates of speech and non-verbal communication

Verbal communication in humans does not come isolated from non-verbal signs; in order to achieve even the most basic degree of naturalness, any humanoid robot needs for example at least some lip-movement-like feature to accompany speech production. Apart from lip-syncing, many other human motor actions are intertwined with speech and natural language; for example, head nods, deictic gestures, gaze movements, etc. Also, note that the term correlates is somewhat misleading; for example, the gesture channel can be more accurately described as being a complementary channel rather than a channel correlated with or just accompanying speech [144]. Furthermore, we are not interested only in the generation of such actions; but also on their combination, as well as on dialogic/interactional aspects.

Let us start by examining the generation of lip syncing. The first question that arises is: should lip sync actions be generated from phoneme-level information, or is the speech soundtrack adequate? Simpler techniques, rely on the speech soundtrack only; the simplest solution being to utilize only the loudness of the soundtrack, and map directly from loudness to mouth opening. There are many shortcomings in this approach; for example, a nasal "m" usually has large apparent loudness, although in humans it is being produced with a closed mouth. Generally, the resulting lip movements of this method are perceivable unnatural. As an improvement to the above method, one can try to use spectrum matching of the soundtrack to a set of reference sounds, such as at [145,146], or

even better, a linear prediction speech model, such as [147]. Furthermore, apart from the generation of lip movements, their recognition can be quite useful regarding the improvement of speech recognition performance under low signal-to-noise ratio conditions [148]. There is also ample evidence that humans utilize lip information during recognition; a celebrated example is the McGurk effect [149]. The McGurk effect is an instance of so-called multisensory perception phenomena [150], which also include other interesting cases such as the rubber hand illusion [151].

Yet another important aspect of communication that requires non-verbal elements is backchannel signalling, primarily accomplished through head nods that the listener provides as feedback to the speaker while listening, in order to for example signal acknowledgement of understanding and continued attention, so that the speaker can continue to provide more verbal input to the listener. An example of a study on backchannel head nods for the case of human–robot communication is given in [152].

Now, let us move on to gestures. The simplest form of gestures which are also directly relevant to natural language are deictic gestures, pointing towards an object and usually accompanied with indexicals such as "this one!". Such gestures have long been utilized in human–robot interaction; starting from virtual avatar systems such as Kris Thorissons Gandalf [153], and continuing all the way to robots such as ACE (Autonomous City Explorer) [154], a robot that was able to navigate through Munich by asking pedestrians for directions. There exist quite a number of other types of gestures, depending on the taxonomy one adopts; such as iconic gestures, symbolic gestures, etc. Furthermore, gestures are highly important towards teaching and learning in humans [155]. Apart from McNeills seminal psychological work [144], a definitive reference to gestures, communication, and their relation to language, albeit regarding virtual avatar Embodied Conversational Assistants (ECA), can be found in the work of Justine Cassell, including [156,157]. Many open questions exist in this area; for example, regarding the synchronization between speech and the different non-verbal cues [158], and socio-pragmatic influences on the non-verbal repertoire.

Another important topic for human–robot interaction is eye gaze coordination and shared attention. Eye gaze cues are important for coordinating collaborative tasks [159,160], and also, eye gazes are an important subset of non-verbal communication cues that can increase efficiency and robustness in human–robot teamwork [161]. Furthermore, in the tour guide setting of [42] a robot that engages visitors in mutual gaze is seen as more humanlike, and slight gaze preference biases towards one of the visitors can positively influence attitudes towards the robot. An attention control system for social robots that can adaptively attract and control a target persons attention is described in [162], extending from the work reported in [163]. The design of robot eyes in order to maximize suitability for gaze reading in investigated in [164]. Also, it is worth noting that gaze plays a significant role in tasks such as robot-to-human handovers [165,166].

Furthermore, eye gaze is very important in disambiguating referring expressions, without the need for hand deixis [167,168], and in shaping participant roles in conversation [169]. Shared attention mechanisms develop in humans during infancy [170], and Scasellati authored the pioneering work on shared attention in robots in 1996 [171], followed up by [172]. A developmental viewpoint is also taken in [173], as well as in [174]. A well-cited probabilistic model of gaze imitation and shared attention is given in [175], In virtual avatars, considerable work has also taken place; such as [176,177].

Eye-gaze observations are also very important towards mind reading and theory of mind [178] for robots; i.e. being able to create models of the mental content and mental functions of other agents (human or robots) minds through observation. Children develop a

progressively more complicated theory of mind during their childhood [179]. Elemental forms of theory of mind are very important also towards purposeful speech generation; for example, in creating referring expressions, one should ideally take into account the second-order beliefs of his conversational partner-listener; i.e. he should use his beliefs regarding what he thinks the other person believes, in order to create a referring expression that can be resolved uniquely by his listener. Furthermore, when a robot is purposefully issuing an inform statement ("there is a tomato behind you") it should know that the human does not already know that; i.e. again an estimated model of second-order beliefs is required (i.e. what the robot believes the human believes). A pioneering work in theory of mind for robots is Scasellatis [180,181]. An early implementation of perspective-shifting synthetic-camera-driven second-order belief estimation for the Ripley robot is given in [53]. Another example of perspective shifting with geometric reasoning for the HRP-2 humanoid is given in [182].

A major technoeconomic obstacle in the past decade regarding the widespread use of systems which can monitor and react to human gaze, and estimate human attention, has been the cost of precise wearable eye-tracking systems, and the need of placing artificial landmarks in the field of view in most of the traditional realizations of such systems. However, with recent developments, including the google glasses, the situation is rapidly changing. An innovative method for the estimation of human fixations in 3D environments that does not require artificial landmarks and enables attention mapping in 3D models with high precision, is presented in [183], which might well be promising to bring forth new opportunities for studies in joint attention as well as applications for human–robot interaction.

Finally, a quick note on a related field, which is recently growing. Children with Autistic Spectrum Disorders (ASD) face special communication challenges. A prominent theory regarding autism is hypothesizing theory-of-mind deficiencies for autistic individuals [184,185]. However, recent research [186–189] has indicated that specially-designed robots that interact with autistic children could potentially help them towards improving their communication skills, and potentially transferring over these skills to communicating not only with robots, but also with other humans.

Yet another important observation to be made is concerned with the relation between the human–human and human–robot interactions. Models arising from observing human–human interactions, can later be used as a basis in order to develop and further refine human–robot interaction. An example of a study of human non-verbal behaviours during teaching, which was made with this purpose in mind, is given in [190]. Last but not least, regarding a wider overview of existing work on non-verbal communication between humans, which could readily provide ideas for future human–robot experiments, the interested reader is referred to [24].

3.7. Purposeful speech and planning

Traditionally, simple command-only canned-response conversational robots had dialogue systems that could be construed as stimulus–response tables: a set of verbs or command utterances were the stimuli, the responses being motor actions, with a fixed mapping between stimuli and responses. Even much more advanced systems, that can support situated language, multiple speech acts, and perspective-shifting theory-of-mind, such as Ripley [53], can be construed as effectively being (stimulus, state) to response maps, where the state of the system includes the contents of the situation model of the robots. What is missing in all of these systems is an explicit modelling of purposeful behaviour towards goals.

Since the early days of AI, automated planning algorithms such as the classic STRIPS [191] and purposeful action selection techniques have been a core research topic. In traditional non-embodied dialogue systems practice, approaches such as Belief-Desire-Intention (BDI) have existed for a while [192], and theoretical models for purposeful generation of speech acts [193] and computation models towards speech planning exist since more than two decades. Also, in robotics, specialized modified planning algorithms have mainly been applied towards motor action planning and path planning [191], such as RRT [194] and Fast-Marching Squares [195].

However, it is worth mentioning that the traditional approach towards planning and reasoning faces a very important problem when applied in real-world robots: a considerable amount of uncertainty exists, arising from the imperfections of the current state-of-the-art of speech and language processing as applied on robots, as well as from the multiple sources and often considerable variance of robot sensory-motor errors. Thus, special techniques are required, supporting graceful operation under considerable uncertainty: and one of the dominant mathematical approaches towards this problem involves Partially Observable Markov Decision Processes (POMDPs). Kaelbling, Littman, and Cassandra [196] for the case of robot planning, and Young [197] have proposed the usage of such POMPDP models. Also, methods for representation and reasoning with probabilistic knowledge, such as those described in [198], can play an important role towards dealing with uncertainty. Thus, such tools provide interesting avenues for wider application in real-world interactive robots, and show a highly promising direction for bridging the gap between symbolic representations and the noisy sensorymotor data of real-world robots.

However, the important point to notice here is that, although considerable research exists for motor planning or dialogue planning alone, there are almost no systems and generic frameworks either for effectively combining the two, or for having mixed speech- and motor-act planning, or even better agent- and object-interaction-directed planners. Notice that motor planning and speech planning cannot be isolated from one another in real-world systems; both types of actions are often interchangeable with one another towards achieving goals, and thus should not be planned by separate subsystems which are independent of one another. For example, if a robot wants to lower its temperature, it could either say: can you kindly open the window? to a human partner (speech action), or could move its body, approach the window, and close it (motor action). An exemption to this research void of mixed speech-motor planning is [199], where a basic purposeful action selection system for question generation or active sensing act generation is described, and implemented on a real conversation robot. However, this is an early and quite task-specific system, and thus much more remains to be done towards real-world general mixed speech act and motor act action selection and planning for robots.

3.8. Multi-level learning

Yet another challenge towards fluid verbal and non-verbal human–robot communication is concerned with learning [200]. But when could learning take place, and what could be and should be learnt? Let us start by examining the when. Data-driven learning can happen at various stages of the lifetime of a system: it could either take place (a) initially and offline, at design time; or, it could take place (b) during special learning sessions, where specific aspects and parameters of the system are renewed; or, (c) it could take place during normal operation of the system, in either a human-directed manner, or ideally (d) through robot-initiated active learning during normal operation. Most current systems that exhibit learning, are actually involving offline learning, i.e. case (a) from above. No systems in the literature have exhibited non-trivial online, real-world continuous learning of communications abilities.

The second aspect beyond the when, is the what of learning. What could be ideally, what could be practically, and what should be learnt, instead of pre-coded, when it comes to human–robot communication? For example, when it comes to natural-language communication, multiple layers exist: the phonological, the morphological, the syntactic, the semantic, the pragmatic, the dialogic. And if one adds the complexity of having to address the symbol grounding problem, a robot needs to have models of grounded meaning, too, in a certain target space, for example in a sensory-motor or a teleological target space. This was already discussed in the previous sections of normative vs. empirical meaning and on symbol grounding. Furthermore, such models might need to be adjustable on the fly; as discussed in the section on online negotiation of meaning. Also, many different aspects of non-verbal communication, from facial expressions to gestures to turn-taking, could ideally be learnable in real operation, even more so for the future case of robots needing to adapt to cultural and individual variations in non-verbal communications. Regarding motor aspects of such non-verbal cues, existing methods in imitation and demonstration learning [28] have been and could further be readily adapted; see for example the imitation learning of human facial expressions for the Leonardo robot [201].

Finally, another important caveat needs to be spelt out at this point. Real-world learning and real-world data collection towards communicative behaviour learning for robots, depending on the data set size required, might require many hours of uninterrupted operation daily by numerous robots: a requirement which is quite unrealistic for today's systems. Therefore, other avenues need to be sought towards acquiring such data sets; and crowdsourcing through specially designed online games offers a realistic potential solution, as mentioned in the previous paragraph on real-world acquisition of large-scale models of grounding. And of course, the learning content of such systems can move beyond grounded meaning models, to a wider range of the what that could be potentially learnable. A relevant example from a non-embodied setting comes from [202], where a chatterbot acquired interaction capabilities through massive observation and interaction with humans in chat rooms. Of course, there do exist inherent limitations in such online systems, even for the case of the robot-tailored online games such as [107]; for example, the non-physicality of the interaction presents specific obstacles and biases. Being able to extend this promising avenue towards wider massive data-driven models, and to demonstrate massive transfer of learning from the online systems to real-world physical robots, is thus an important research avenue for the future.

3.9. Utilization of online resources and services

Yet another interesting avenue towards enhanced human–robot communication that has opened up recently is the following: as more and more robots nowadays can be constantly connected to the internet, not all data and programs that the robot uses need to be onboard its hardware. Therefore, a robot could potentially utilize online information as well as online services, in order to enhance its communication abilities. Thus, the intelligence of the robot is partially offloaded to the internet; and potentially, thousands of programs and/or humans could be providing part of its intelligence, even in real-time. For example, going much beyond traditional cloud robotics [203], in the human–robot cloud proposal [204], one could construct on-demand and on-the-fly distributed robots with human and machine sensing, actuation, and processing components.

Beyond these highly promising glimpses of a possible future, there exist a number of implemented systems that utilize information and/or services from the internet. A prime example is Facebots, which are physical robots that utilize and publish information

on Facebook towards enhancing long-term human–robot interaction, are described in [66,67]. Facebots are creating shared memories and shared friends with both their physical as well as their online interaction partners, and are utilizing this information towards creating dialogues that enable the creation of a longer-lasting relationship between the robot and its human partners, thus reversing the quick withdrawal of the novelty effects of long-term HRI reported in [205]. Also, as reported in [206], the multi-lingual conversational robot Ibn Sina [48], has made use of online google translate services, as well as wikipedia information for its dialogues. Furthermore, one could readily utilize online high-quality speech recognition and text-to-speech services for human–robot communication, such as [Sonic Cloud online services], in order not to sacrifice onboard computational resources.

Also, quite importantly, there exists the European project Roboearth [207], which is described as a World Wide Web for robots: a giant network and database repository where robots can share information and learn from each other about their behaviour and their environment. Bringing a new meaning to the phrase experience is the best teacher, the goal of RoboEarth is to allow robotic systems to benefit from the experience of other robots, paving the way for rapid advances in machine cognition and behaviour, and ultimately, for more subtle and sophisticated human–machine interaction. Rapyuta [208], which is the cloud engine of Roboearth, claims to make immense computational power available to robots connected to it. Of course, beyond what has been utilized so far, there are many other possible sources of information and/or services on the internet to be exploited; and thus much more remains to be done in the near future in this direction.

3.10. Miscellaneous abilities

Beyond the nine desiderata examined so far, there exist a number of other abilities that are required towards fluid and general human–robot communication. These have to do with dealing with multiple conversational partners in a discussion, with support for multilingual capabilities, and with generating and recognizing natural language across multiple modalities: for example not only acoustic, but also in written form. In more detail:

3.10.1. Multiple conversational partners

Regarding conversational turn-taking, in the words of Sacks [209], the organization of taking turns to talk is fundamental to conversation, as well as to other speech-exchange systems, and this readily carries over to human–robot conversations, and becomes especially important in the case of dialogues with multiple conversation partners. Recognition of overlapping speech is also quite important towards turn-taking [210]. Regarding turn-taking in robots, a computational strategy for robots participating in group conversation is presented in [211], and the very important role of gaze cues in turn taking and participant role assignment in human–robot conversations is examined in [169]. In [212], an experimental study using the robot Simon is reported, which is aiming towards showing that the implementation of certain turn-taking cues can make interaction with a robot easier and more efficient for humans. Head movements are also very important in turn-taking; the role of which in keeping engagement in an interaction is explored in [213].

Yet another requirement for fluid multi-partner conversations is sound-source localization and speaker identification. Sound source localization is usually accomplished using microphone arrays, such as the robotic system in [214]. An approach utilizing scattering theory for sound source localization in robots is described in [215] and approaches using beamforming for multiple moving sources are presented in [216,217]. Finally, HARK, an

open-source robot audition system supporting three simultaneous speakers, is presented in [218]. Speaker identification is an old problem; classic approaches utilize Gaussian mixture models, such as [219,220]. Robotic systems able to identify their speakers identity include [221,64], as well as the well-cited [222]. Also, an important idea towards effective signal separation between multiple speaker sources in order to aid in recognition, is to utilize both visual as well as auditory information towards that goal. Classic examples of such approaches include [223], as well as [224].

Of course, special considerations are needed not only in the case of multiple verbal conversational partners, but also multiple interactive communication partners at large, also covering non-verbal components. An example of an assisted care robot serving tea to the elderly which is able to gracefully deal with requests from multiple individuals simultaneously is presented in [47].

3.10.2. Multilingual capabilities and mutimodal natural language

Yet another desirable ability for human–robot communication is multilinguality. Multilingual robots could not only communicate with a wider range of people, especially in multi-cultural societies and settings such as museums, but could very importantly also act as translators and mediators. Although there has been considerable progress towards non-embodied multilingual dialogue systems [225], and multi-lingual virtual avatars do exist [226,227], the only implemented real-world multilingual physical android robot so far reported in the literature is [206].

Finally, let us move on to examining multiple modalities for the generation and recognition of natural language. Apart from a wealth of existing research on automated production and recognition of sign language for the deaf (ASL) [228–230], systems directly adaptable to robots also exist [231]. One could also investigate the intersection between human writing and robotics. Again, a wealth of approaches exist for the problem of optical character recognition and handwriting recognition [232,233], even for languages such as Arabic [234], the only robotic system that has demonstrated limited OCR capabilities is [206]. Last but not least, another modality available for natural language communication for robots is internet chat. The only reported system so far that could perform dialogues both physically as well as through facebook chat is [66,67].

As a big part of human knowledge, information, as well as real-world communication is taking place either through writing or through such electronic channels, inevitably more and more systems in the future will have corresponding abilities. Thus, robots will be able to more fluidly integrate within human societies and environments, and ideally will be enabled to utilize the services offered within such networks for humans. Most importantly, robots might also one day become able to help maintain and improve the physical human–robot social networks they reside within towards the benefit of the common good of all, as is advocated in [235].

4. Discussion

From our detailed examination of the ten desiderata, what follows first is that although we have moved beyond the canned-commands-only, canned responses state-of-affairs of the nineties, we seem to be still far from our goal of fluid and natural verbal and non-verbal communication between humans and robots. But what is missing?

Many promising future directions were mentioned in the preceding sections. Apart from clearly open avenues for projects in a number of areas, such as composition of grounded semantics, online negotiation of meaning, affective interaction and closed-loop affective dialogue, mixed speech-motor planning, massive acquisition of data-driven models for human–robot communication through crowd-sourced online games, real-time exploitation of online information and services for enhanced human–robot communication, many more open areas exist.

What we speculate might really make a difference, though, is the availability of massive real-world data, in order to drive further data-driven models. And in order to reach that state, a number of robots need to start getting deployed, even if in partially autonomous partially remote-human-operated mode, in real-world interactive application settings with round-the-clock operation: be it shopping mall assistants, receptionists, museum robots, or companions, the application domains that will bring out human–robot communication to the world in more massive proportions, remain yet to be discovered. However, given recent developments, it does not seem to be so far away anymore; and thus, in the coming decades, the days might well come when interactive robots will start being part of our everyday lives, in seemless harmonious symbiosis, hopefully helping to create a better and exciting future.

5. Conclusions

An overview of research in human–robot interactive communication was presented, covering verbal as well as non-verbal aspects. Following a historical introduction reaching from roots in antiquity to well into the nineties, and motivation towards fluid human–robot communication, ten desiderata were proposed, which provided an organizational axis both of recent as well as of future research on human–robot communication. Then, the ten desiderata were explained, relevant research was examined in detail, culminating to a unifying discussion. In conclusion, although almost twenty-five years in human–robot interactive communication exist, and significant progress has been achieved in many fronts, many sub-problems towards fluid verbal and non-verbal human–robot communication remain yet unsolved, and present highly promising and exciting avenues towards research in the near future.

References

[1] L. Ballard, Robotics'founding father George C. Devol-serial entrepreneur and inventor, Robot-Congers (31) (2011) 58.
[2] G.C. Devol, Encoding apparatus, January 24 1984, US Patent 4,427,970.
[3] D.-L. Gera, Ancient Greek Ideas on Speech, Language, and Civilization, Oxford University Press, 2003.
[4] R. Lattimore, R. Martin, The Iliad of Homer, University of Chicago Press, 2011.
[5] C. Huffman, Archytas of Tarentum: Pythagorean, Philosopher and Mathematician King, Cambridge University Press, 2005.
[6] J. Needham, Science and Civilisation in China: Volume 2, Cambridge University Press, 1959.
[7] N. Sharkey, The programmable robot of ancient Greece, New Scientist, July 2007, pp. 32–35.
[8] M.E. Rosheim, Robot Evolution: The Development of Anthrobotics, first ed., John Wiley & Sons, Inc., New York, NY, USA, 1994.
[9] N. Hockstein, C. Gourin, R. Faust, D. Terris, A history of robots: from science fiction to surgical robotics, J. Robot. Surg. 1 (2) (2007) 113–118.
[10] D.H. Klatt, Review of text-to-speech conversion for English, J. Acoust. Soc. Am. 82 (3) (1987) 737–793.
[11] G. Antoniol, R. Cattoni, M. Cettolo, M. Federico, Robust speech understanding for robot telecontrol, in: Proceedings of the 6th International Conference on Advanced Robotics, 1993, pp. 205–209.
[12] W. Burgard, A.B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thrun, The interactive museum tour-guide robot, in: Proc. of the Fifteenth National Conference on Artificial Intelligence, AAAI-98, 1998.
[13] L. Versweyveld, Voice-controlled surgical robot ready to assist in minimally invasive heart surgery, Virtual Medicine World Monthly, March 1998.
[14] I. Horswill, Polly: a vision-based artificial agent, in: Proceedings of the Eleventh National Conference on Artificial Intelligence, AAAI-93, Press, 1993, pp. 824–829.
[15] I. Horswill, The design of the Polly system, Tech. Rep., The Institute for the Learning Sciences, Northwestern University, 1996, September.
[16] M. Torrance, Natural communication with mobile robots (Master's thesis), MIT Department of Electrical Engineering and Computer Science, 1994, January.
[17] G. Antoniol, B. Caprile, A. Cimatti, R. Fiutem, Experiencing real-life interactions with the experimental platform of MAIA, in: Proceedings of the 1st European Workshop on Human Comfort and Security, 1994.

[18] I. Androutsopoulos, A principled framework for constructing natural language interfaces to temporal databases (Ph.D. dissertation), Department of Artificial Intelligence, University of Edinburgh, 1996.

[19] H. Asoh, T. Matsui, J. Fry, F. Asano, S. Hayamizu, A spoken dialog system for a mobile office robot, in: Proceedings of the European Conference on Speech Communication and Technology, EUROSPEECH, ISCA, 1999.

[20] J. Fry, H. Asoh, T. Matsui, Natural dialogue with the Jijo-2 office robot, in: 1998 IEEE/RSJ International Conference on Intelligent Robots and Systems, Proceedings, vol. 2, 1998, pp. 1278–1283.

[21] T. Matsui, H. Asoh, J. Fry, Y. Motomura, F. Asano, T. Kurita, I. Hara, N. Otsu, Integrated natural spoken dialogue system of Jijo-2 mobile robot for office services, in: Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, ser. AAAI'99/IAAI'99, American Association for Artificial Intelligence, Menlo Park, CA, USA, 1999, pp. 621–627.

[22] C. Crangle, P. Suppes, Language and Learning for Robots, in: Ser. CSLI Lecture Notes, Center for the Study of Language and Information, 1994, [Online]. Available: http://books.google.gr/books?id=MlMQ11Pqz10C.

[23] J.R. Searle, Speech Acts: An Essay in the Philosophy of Language, Cambridge University Press, Cambridge, 1969.

[24] K. Vogeley, G. Bente, "Artificial humans": psychology and neuroscience perspectives on embodiment and nonverbal communication, Neural Netw. 23 (8) (2010) 1077–1090.

[25] S. Harnad, The symbol grounding problem, Physica D 42 (1) (1990) 335–346.

[26] G. Schreiber, A. Stemmer, R. Bischoff, The fast research interface for the KUKA lightweight robot, in: IEEE Conference on Robotics and Automation, ICRA, 2010.

[27] S. Wrede, C. Emmerich, R. Grünberg, A. Nordmann, A. Swadzba, J. Steil, A user study on kinesthetic teaching of redundant robots in task and configuration space, J. Hum.–Robot Interact. 2 (1) (2013) 56–81.

[28] B.D. Argall, S. Chernova, M. Veloso, B. Browning, A survey of robot learning from demonstration, Robot. Auton. Syst. 57 (5) (2009) 469–483.

[29] C.L. Nehaniv, K. Dautenhahn, Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions, Cambridge University Press, 2007.

[30] N. Mavridis, D. Roy, Grounded situation models for robots: where words and percepts meet, in: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006, pp. 4690–4697.

[31] T. van der Zant, T. Wisspeintner, RoboCup@ home: creating and benchmarking tomorrows service robot applications, Robotic Soccer, 2007, pp. 521–528.

[32] M.E. Foster, T. By, M. Rickert, A. Knoll, Human–robot dialogue for joint construction tasks, in: Proceedings of the 8th International Conference on Multimodal Interfaces, ser. ICMI'06, ACM, New York, NY, USA, 2006, pp. 68–71.

[33] M. Giuliani, A. Knoll, Evaluating supportive and instructive robot roles in human–robot interaction, in: Social Robotics, Springer, 2011, pp. 193–203.

[34] K. Wada, T. Shibata, Living with seal robots—its sociopsychological and physiological influences on the elderly at a care house, IEEE Trans. Robot. 23 (5) (2007) 972–980.

[35] K. Kamei, K. Shinozawa, T. Ikeda, A. Utsumi, T. Miyashita, N. Hagita, Recommendation from robots in a real-world retail shop, in: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ACM, 2010, p. 19.

[36] V. Chidambaram, Y.-H. Chiang, B. Mutlu, Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues, in: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'12, ACM, New York, NY, USA, 2012, pp. 293–300. [Online]. Available: http://doi.acm.org/10.1145/2157689.2157798.

[37] M. Makatchev, I. Fanaswala, A. Abdulsalam, B. Browning, W. Ghazzawi, M. Sakr, R. Simmons, Dialogue patterns of an arabic robot receptionist, in: 2010 5th ACM/IEEE International Conference on Human–Robot Interaction, HRI, 2010, pp. 167–168.

[38] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, N. Hagita, An affective guide robot in a shopping mall, in: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, ser. HRI'09, ACM, New York, NY, USA, 2009, pp. 173–180. [Online]. Available: http://doi.acm.org/10.1145/1514095.1514127.

[39] M. Shiomi, T. Kanda, H. Ishiguro, N. Hagita, Interactive humanoid robots for a science museum, in: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human–Robot Interaction, ser. HRI'06, ACM, New York, NY, USA, 2006, pp. 305–312. [Online]. Available: http://doi.acm.org/10.1145/1121241.1121293.

[40] A. Yamazaki, K. Yamazaki, T. Ohyama, Y. Kobayashi, Y. Kuno, A techno-sociological solution for designing a museum guide robot: regarding choosing an appropriate visitor, in: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'12, ACM, New York, NY, USA, 2012, pp. 309–316. [Online]. Available: http://doi.acm.org/10.1145/2157689.2157800.

[41] V. Evers, N. Menezes, L. Merino, D. Gavrila, F. Nabais, M. Pantic, P. Alvito, D. Karreman, The development and real-world deployment of frog, the fun robotic outdoor guide, in: Proceedings of the 2014 ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'14, ACM, New York, NY, USA, 2014, p. 100. [Online]. Available: http://doi.acm.org/10.1145/2559636.2559649.

[42] D.E. Karreman, G.U. Sepúlveda Bradford, B.E. van Dijk, M. Lohse, V. Evers, What happens when a robot favors someone? How a tour guide robot uses gaze behavior to address multiple persons while storytelling about art, in: Proceedings of the 8th ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'13, IEEE Press, Piscataway, NJ, USA, 2013, pp. 157–158. [Online]. Available: http://dl.acm.org/citation.cfm?id=2447556.2447615.

[43] V. Evers, R. de Vries, P. Alvito, Designing interruptive behaviors of a public environmental monitoring robot, in: Proceedings of the 6th International Conference on Human–Robot Interaction, ser. HRI'11, ACM, New York, NY, USA, 2011, pp. 131–132. [Online]. Available: http://doi.acm.org/10.1145/1957656.1957696.

[44] S. Tellex, D. Roy, Spatial routines for a simulated speech-controlled vehicle, in: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human–Robot Interaction, ACM, 2006, pp. 156–163.

[45] Y. Kobayashi, Y. Kinpara, E. Takano, Y. Kuno, K. Yamazaki, A. Yamazaki, A wheelchair which can automatically move alongside a caregiver, in: Proceedings of the 6th International Conference on Human–Robot Interaction, ser. HRI'11, ACM, New York, NY, USA, 2011, pp. 407–408. [Online]. Available: http://doi.acm.org/10.1145/1957656.1957805.

[46] K. Dautenhahn, M. Walters, S. Woods, K.L. Koay, C.L. Nehaniv, A. Sisbot, R. Alami, T. Siméon, How may i serve you? A robot companion approaching a seated person in a helping context, in: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human–Robot Interaction, ACM, 2006, pp. 172–179.

[47] Y. Kobayashi, M. Gyoda, T. Tabata, Y. Kuno, K. Yamazaki, M. Shibuya, Y. Seki, Assisted-care robot dealing with multiple requests in multi-party settings, in: Proceedings of the 6th International Conference on Human–Robot Interaction, ser. HRI'11, ACM, New York, NY, USA, 2011, pp. 167–168. [Online]. Available: http://doi.acm.org/10.1145/1957656.1957714.

[48] N. Mavridis, D. Hanson, The IbnSina Center: an augmented reality theater with intelligent robotic and virtual characters, in: The 18th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2009, IEEE, 2009, pp. 681–686.

[49] J. Zlotowski, T. Bleeker, C. Bartneck, R. Reynolds, I sing the body electric: an experimental theatre play with robots, in: Proceedings of the 8th ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'13, IEEE Press, Piscataway, NJ, USA, 2013, pp. 427–428. [Online]. Available: http://dl.acm.org/citation.cfm?id=2447556.2447702.

[50] K. Petersen, J. Solis, A. Takanishi, Musical-based interaction system for the waseda flutist robot, Auton. Robots 28 (4) (2010) 471–488.

[51] K. Kosuge, T. Hayashi, Y. Hirata, R. Tobiyama, Dance partner robot-MS Dancer, in: 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2003, Proceedings, vol. 4, IEEE, 2003, pp. 3459–3464.

[52] V.A. Kulyukin, On natural language dialogue with assistive robots, in: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human–Robot Interaction, ACM, 2006, pp. 164–171.

[53] N. Mavridis, Grounded situation models for situated conversational assistants (Ph.D. dissertation), Massachusetts Institute of Technology, 2007.

[54] T. Winograd, A procedural model of language understanding, in: Readings in Natural Language Processing, Morgan Kaufmann Publishers Inc., 1986, pp. 249–266.

[55] T. Winograd, Understanding natural language, Cogn. Psychol. 3 (1) (1972) 1–191.

[56] J. Dzifcak, M. Scheutz, C. Baral, P. Schermerhorn, What to do and how to do it: translating natural language directives into temporal and dynamic logic representation for goal management and action execution, in: Proceedings of the 2009 IEEE International Conference on Robotics and Automation, ICRA'09, Kobe, Japan, May 2009.

[57] J. Austin, How to Do Things with Words. Oxford, 1962.

[58] J. Searle, A taxonomy of illocutionary acts, in: K. Gunderson (Ed.), Language, Mind and Knowledge, University of Minnesota Press, 1975, pp. 344–369.

[59] J.F. Allen, D.K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, A. Stent, Towards conversational human–computer interaction, AI Mag. 22 (2001) 27–37.

[60] P.N. Johnson-Laird, Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness, Vol. 6, Harvard University Press, 1983.

[61] R.A. Zwaan, G.A. Radvansky, Situation models in language comprehension and memory, Psychol. Bull. 123 (2) (1998) 162.

[62] H.P. Grice, Logic and conversation, 1975, pp. 41–58.

[63] S. Wilske, G.-J. Kruijff, Service robots dealing with indirect speech acts, in: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2006, pp. 4698–4703.

[64] F. Krsmanovic, C. Spencer, D. Jurafsky, A.Y. Ng, Have we met? MDP based speaker ID for robot dialogue. in: INTERSPEECH, 2006.

[65] C.T. Ishi, H. Ishiguro, N. Hagita, Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts. in: INTERSPEECH, 2006.

[66] N. Mavridis, M. Petychakis, A. Tsamakos, P. Toulis, S. Emami, W. Kazmi, C. Datta, C. BenAbdelkader, A. Tanoto, FaceBots: steps towards enhanced long-term human–robot interaction by utilizing and publishing online social information, Paladyn 1 (3) (2010) 169–178.

[67] N. Mavridis, C. Datta, S. Emami, A. Tanoto, C. BenAbdelkader, T. Rabie, FaceBots: robots utilizing and publishing social information in facebook, in: 2009 4th ACM/IEEE International Conference on Human–Robot Interaction, HRI, IEEE, 2009, pp. 273–274.

[68] B. Wrede, S. Buschkaemper, C. Muhl, K.J. Rohlfing, Analyses of feedback in HRI, in: How People Talk to Computers, Robots, and Other Artificial Communication Partners, 2006, p. 38.

[69] R. Stiefelhagen, H.K. Ekenel, C. Fugen, P. Gieselmann, H. Holzapfel, F. Kraft, K. Nickel, M. Voit, A. Waibel, Enabling multimodal human–robot interaction for the Karlsruhe humanoid robot, IEEE Trans. Robot. 23 (5) (2007) 840–851.

[70] D. Ertl, A. Green, H. Hüttenrauch, F. Lerasle, Improving human–robot communication with mixed-initiative and context-awareness co-located with Ro-Man 2009.

[71] M. Ralph, M.A. Moussa, Toward a natural language interface for transferring grasping skills to robots, IEEE Trans. Robot. 24 (2) (2008) 468–475.

[72] J. Weizenbaum, Eliza—a computer program for the study of natural language communication between man and machine, Commun. ACM 9 (1) (1966) 36–45.

[73] M.L. Mauldin, Chatterbots, tinymuds, and the Turing test: entering the Loebner prize competition, in: AAAI, vol. 94, 1994, pp. 16–21.

[74] D. Roy, A computational model of word learning from multimodal sensory input, in: Proceedings of the International Conference on Cognitive Modeling, ICCM2000, Groningen, Netherlands. Citeseer, 2000.

[75] R. Baillargeon, E.S. Spelke, S. Wasserman, Object permanence in five-month-old infants, Cognition 20 (3) (1985) 191–208.

[76] D. Roy, K.-Y. Hsiao, N. Mavridis, Mental imagery for a conversational robot, IEEE Trans. Syst. Man Cybern. B 34 (3) (2004) 1374–1383.

[77] E. Tulving, Elements of Episodic Memory, Clarendon Press Oxford, 1983.

[78] N. Mavridis, M. Petychakis, Human-like memory systems for interactive robots: desiderata and two case studies utilizing groundedsituation models and online social networking.

[79] D. Gentner, K.D. Forbus, Computational models of analogy, Wiley Interdiscip. Rev.: Cogn. Sci. 2 (3) (2011) 266–276.

[80] C.S. Pierce, Logic as semiotic: the theory of signs, in: The Philosophical Writings of Pierce, 1955, pp. 98–119.

[81] C.S. Peirce, Collected Papers of Charles Sanders Peirce, Vol. 3, Harvard University Press, 1974.

[82] T. Spexard, S. Li, B. Wrede, J. Fritsch, G. Sagerer, O. Booij, Z. Zivkovic, B. Terwijn, B. Krose, BIRON, where are you? Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization, in: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2006, pp. 934–940.

[83] L. Steels, Evolving grounded communication for robots, Trends Cogn. Sci. 7 (7) (2003) 308–312.

[84] S.D. Larson, Intrinsic Representation: Bootstrapping Symbols from Experience, Springer, 2004.

[85] P.J. Gorniak, The affordance-based concept (Ph.D. dissertation), Massachusetts Institute of Technology, 2005.

[86] C. Yu, L.B. Smith, A.F. Pereira, Grounding word learning in multimodal sensorimotor interaction, in: Proceedings of the 30th Annual Conference of the Cognitive Science Society, 2008, pp. 1017–1022.

[87] T. Regier, L.A. Carlson, Grounding spatial language in perception: an empirical and computational investigation, J. Exp. Psychol. [Gen.] 130 (2) (2001) 273.

[88] D.K. Roy, Learning visually grounded words and syntax for a scene description task, Comput. Speech Lang. 16 (3) (2002) 353–385.

[89] K.R. Coventry, S.C. Garrod, Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions, Psychology Press, 2004.

[90] J.M. Siskind, Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic, 2011. ArXiv Preprint arXiv:1106.0256.

[91] J. Zlatev, Spatial semantics, in: Handbook of Cognitive Linguistics, 2007, pp. 318–350.

[92] M. Skubic, D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, D. Brock, Spatial language for human–robot dialogs, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 34 (2) (2004) 154–167.

[93] H. Zender, O. Martínez Mozos, P. Jensfelt, G.-J. Kruijff, W. Burgard, Conceptual spatial representations for indoor mobile robots, Robot. Auton. Syst. 56 (6) (2008) 493–502.

[94] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A.G. Banerjee, S.J. Teller, N. Roy, Understanding natural language commands for robotic navigation and mobile manipulation, in: AAAI, 2011.

[95] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A.G. Banerjee, S. Teller, N. Roy, Approaching the symbol grounding problem with probabilistic graphical models, AI Mag. 32 (4) (2011) 64–76.

[96] S. Tellex, P. Thaker, J. Joseph, N. Roy, Learning perceptually grounded word meanings from unaligned parallel data, Mach. Learn. (2013) 1–17.

[97] K. Gold, B. Scassellati, Grounded pronoun learning and pronoun reversal, in: Proceedings of the 5th International Conference on Development and Learning, 2006.

[98] L. Steels, The symbol grounding problem has been solved. so what's next, in: Symbols and Embodiment: Debates on Meaning and Cognition, 2008, pp. 223–244.

[99] L. Steels, Semiotic dynamics for embodied agents, IEEE Intell. Syst. 21 (3) (2006) 32–38.

[100] C. Hudelot, N. Maillot, M. Thonnat, Symbol grounding for semantic image interpretation: from image data to semantics, in: Tenth IEEE International Conference on Computer Vision Workshops, ICCVW'05, IEEE, 2005.

[101] A.M. Cregan, Symbol grounding for the semantic Web, in: The Semantic Web: Research and Applications, Springer, 2007, pp. 429–442.

[102] C. Wallraven, M. Schultze, B. Mohler, A. Vatakis, K. Pastra, The poeticon enacted scenario corpus—a tool for human and computational experiments on action understanding, in: 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops, FG 2011, IEEE, 2011, pp. 484–491.

[103] K. Pastra, C. Wallraven, M. Schultze, A. Vataki, K. Kaulard, The poeticon corpus: capturing language use and sensorimotor experience in everyday interaction. in: LREC, Citeseer, 2010.

[104] P. Gärdenfors, Conceptual Spaces: The Geometry of Throught, MIT press, 2004.

[105] J. Orkin, D. Roy, The restaurant game: learning social behavior and language from thousands of players online, J. Game Dev. 3 (1) (2007) 39–60.

[106] S. Chernova, J. Orkin, C. Breazeal, Crowdsourcing HRI through online multi-player games, in: Proc. Dialog with Robots: AAAI Fall Symposium, 2010.

[107] N. DePalma, S. Chernova, C. Breazeal, Leveraging online virtual agents to crowdsource human–robot interaction, in: Proceedings of CHI Workshop on Crowdsourcing and Human Computation, 2011.

[108] L. Von Ahn, R. Liu, M. Blum, Peekaboom: a game for locating objects in images, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2006, pp. 55–64.

[109] E.R. Hilgard, The trilogy of mind: cognition, affection, and conation, J. Hist. Behav. Sci. 16 (2) (1980) 107–117.

[110] R.W. Picard, Affective computing: challenges, Int. J. Hum.–Comput. Stud. 59 (1) (2003) 55–64.

[111] R. Picard, S. Papert, W. Bender, B. Blumberg, C. Breazeal, D. Cavallo, T. Machover, M. Resnick, D. Roy, C. Strohecker, Affective learning—a manifesto, BT Technol. J. 22 (4) (2004) 253–269.

[112] G. Haddock, G.R. Maio, K. Arnold, T. Huskinson, Should persuasion be affective or cognitive? The moderating effects of need for affect and need for cognition, Pers. Soc. Psychol. Bull. 34 (6) (2008) 769–778.

[113] C. Breazeal, Emotion and sociable humanoid robots, Int. J. Hum.–Comput. Stud. 59 (1) (2003) 119–155.

[114] J. Cassell, Embodied conversational interface agents, Commun. ACM 43 (4) (2000) 70–78.

[115] W.L. Johnson, J.W. Rickel, J.C. Lester, Animated pedagogical agents: face-to-face interaction in interactive learning environments, Int. J. Artif. Intell. Educ. 11 (1) (2000) 47–78.

[116] F.d. Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, B.D. Carolis, From Greta's mind to her face: modelling the dynamics of affective states in a conversational embodied agent, Int. J. Hum.–Comput. Stud. 59 (1) (2003) 81–118.

[117] C. Breazeal, J. Velásquez, Toward teaching a robot "infant" using emotive communication acts, in: Proceedings of the 1998 Simulated Adaptive Behavior Workshop on Socially Situated Intelligence, 1998, pp. 25–40.

[118] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M.J. Russell, M. Wong, "You stupid tin box"-children interacting with the AIBO robot: a cross-linguistic emotional speech corpus. in: LREC, 2004.

[119] K. Komatani, R. Ito, T. Kawahara, H.G. Okuno, Recognition of emotional states in spoken dialogue with a robot, in: Innovations in Applied Artificial Intelligence, Springer, 2004, pp. 413–423.

[120] B.-C. Bae, A. Brunete, U. Malik, E. Dimara, J. Jermsurawong, N. Mavridis, Towards an empathizing and adaptive storyteller system, in: Eighth Artificial Intelligence and Interactive Digital Entertainment Conference, 2012.

[121] T. Ruf, A. Ernst, C. Küblbeck, Face detection with the sophisticated high-speed object recognition engine (SHORE), in: Microelectronic Systems, Springer, 2011, pp. 243–252.

[122] R. El Kaliouby, P. Robinson, Real-time inference of complex mental states from facial expressions and head gestures, in: Real-Time Vision for Human–Computer Interaction, Springer, 2005, pp. 181–200.

[123] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816.

[124] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Fully automatic facial action recognition in spontaneous behavior, in: 7th International Conference on Automatic Face and Gesture Recognition, FGR 2006, IEEE, 2006, pp. 223–230.

[125] T. Wu, N.J. Butko, P. Ruvulo, M.S. Bartlett, J.R. Movellan, Learning to make facial expressions, in: IEEE 8th International Conference on Development and Learning, ICDL 2009, IEEE, 2009, pp. 1–6.

[126] M.-J. Han, C.-H. Lin, K.-T. Song, Robotic emotional expression generation based on mood transition and personality model, IEEE Trans. Cybern. 43 (4) (2013) 1290–1303.

[127] T. Baltrušaitis, L.D. Riek, P. Robinson, Synthesizing expressions using facial feature point tracking: how emotion is conveyed, in: Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments, ACM, 2010, pp. 27–32.

[128] G. Littlewort, M.S. Bartlett, I. Fasel, J. Susskind, J. Movellan, Dynamics of facial expression extracted automatically from video, Image Vis. Comput. 24 (6) (2006) 615–625.

[129] M. Schröder, Expressive speech synthesis: past, present, and possible futures, in: Affective Information Processing, Springer, 2009, pp. 111–126.

[130] D. Ververidis, C. Kotropoulos, Emotional speech recognition: resources, features, and methods, Speech Commun. 48 (9) (2006) 1162–1181.

[131] S. Roehling, B. MacDonald, C. Watson, Towards expressive speech synthesis in English on a robotic platform, in: Proceedings of the Australasian International Conference on Speech Science and Technology, 2006, pp. 130–135.

[132] A. Chella, R.E. Barone, G. Pilato, R. Sorbello, An emotional storyteller robot. in: AAAI Spring Symposium: Emotion, Personality, and Social Behavior, 2008, pp. 17–22.

[133] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retr. 2 (1–2) (2008) 1–135.

[134] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis, Comput. Linguist. 35 (3) (2009) 399–433.

[135] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, M. Stede, Lexicon-based methods for sentiment analysis, Comput. Linguist. 37 (2) (2011) 267–307.

[136] R.W. Picard, Measuring affect in the wild, in: Affective Computing and Intelligent Interaction, Springer, 2011, p. 3.

[137] S.H. Fairclough, Fundamentals of physiological computing, Interact. Comput. 21 (1) (2009) 133–145.

[138] C.L. Bethel, R.R. Murphy, Affective expression in appearance constrained robots, in: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human–Robot Interaction, ser. HRI'06, ACM, New York, NY, USA, 2006, pp. 327–328. [Online]. Available: http://doi.acm.org/10.1145/1121241.1121299.

[139] C.L. Bethel, R.R. Murphy, Non-facial/non-verbal methods of affective expression as applied to robot-assisted victim assessment, in: Proceedings of the ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'07, ACM, New York, NY, USA, 2007, pp. 287–294. [Online]. Available: http://doi.acm.org/10.1145/1228716.1228755.

[140] D. Arroyo, C. Lucho, S.J. Roncal, F. Cuellar, Daedalus: a sUAV for human–robot interaction, in: Proceedings of the 2014 ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'14, ACM, New York, NY, USA, 2014, pp. 116–117. [Online]. Available: http://doi.acm.org/10.1145/2559636.2563709.

[141] R. Read, T. Belpaeme, How to use non-linguistic utterances to convey emotion in child–robot interaction, in: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'12, ACM, New York, NY, USA, 2012, pp. 219–220. [Online]. Available: http://doi.acm.org/10.1145/2157689.2157764.

[142] R. Read, T. Belpaeme, People interpret robotic non-linguistic utterances categorically, in: Proceedings of the 8th ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'13, IEEE Press, Piscataway, NJ, USA, 2013, pp. 209–210. [Online]. Available: http://dl.acm.org/citation.cfm?id=2447556.2447641.

[143] H.A. Elfenbein, N. Ambady, On the universality and cultural specificity of emotion recognition: a meta-analysis, Psychol. Bull. 128 (2) (2002) 203.

[144] D. McNeill, Hand and Mind: What Gestures Reveal about Thought, University of Chicago Press, 1992.

[145] P. Weil, About face, computergraphic synthesis and manipulation of facial imagery (Ph.D. dissertation), Massachusetts Institute of Technology, 1982.

[146] J. Lewis, P. Purcell, Soft machine: a personable interface, in: Proc. of Graphics Interface, vol. 84. Citeseer, 1984, pp. 223–226.

[147] J.P. Lewis, F.I. Parke, Automated lip-synch and speech synthesis for character animation, ACM SIGCHI Bull. 17 (SI) (1987) 143–147.

[148] C. Bregler, Y. Konig, "Eigenlips" for robust speech recognition, in: 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-94, vol. 2, IEEE, 1994, p. II-669.

[149] H. McGurk, J. MacDonald, Hearing lips and seeing voices, Nature (1976) 746–748.

[150] G.A. Calvert, C. Spence, B.E. Stein, The Handbook of Multisensory Processes, MIT Press, 2004.

[151] M. Tsakiris, P. Haggard, The rubber hand illusion revisited: visuotactile integration and self-attribution, J. Exp. Psychol. [Hum. Percept.] 31 (1) (2005) 80.

[152] N. Segato, A. Krogsager, D.G. Jensen, M. Rehm, The role of physical embodiment of humanoid robot interaction: focusing on backchannel head nods in danish first meeting encounters, in: HCI International 2014-Posters' Extended Abstracts, Springer, 2014, pp. 583–587.

[153] K.R. Thorisson, Communicative humanoids: a computational model of psychosocial dialogue skills (Ph.D. dissertation), Massachusetts Institute of Technology, 1996.

[154] G. Lidoris, F. Rohrmuller, D. Wollherr, M. Buss, The autonomous city explorer (ACE) project—mobile robot navigation in highly populated urban environments, in: IEEE International Conference on Robotics and Automation, ICRA'09, IEEE, 2009, pp. 1416–1422.

[155] W.-M. Roth, Gestures: their role in teaching and learning, Rev. Educ. Res. 71 (3) (2001) 365–392.

[156] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, H. Yan, Embodiment in conversational interfaces: Rea, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 1999, pp. 520–527.

[157] J. Cassell, H.H. Vilhjálmsson, T. Bickmore, BEAT: the behavior expression animation toolkit, in: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, ser. SIGGRAPH'01, ACM, New York, NY, USA, 2001, pp. 477–486.

[158] N. Rossini, Patterns of synchronization of non-verbal cues and speech in ECAs: towards a more "natural" conversational agent, in: Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues, Springer, 2011, pp. 96–103.

[159] S.R. Fussell, R.E. Kraut, J. Siegel, Coordination of communication: effects of shared visual context on collaborative work, in: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, ACM, 2000, pp. 21–30.

[160] S.E. Brennan, X. Chen, C.A. Dickinson, M.B. Neider, G.J. Zelinsky, Coordinating cognition: the costs and benefits of shared gaze during collaborative search, Cognition 106 (3) (2008) 1465–1477.

[161] C. Breazeal, C.D. Kidd, A.L. Thomaz, G. Hoffman, M. Berlin, Effects of nonverbal communication on efficiency and robustness in human–robot teamwork, in: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2005, IEEE, 2005, pp. 708–713.

[162] D. Das, M.M. Hoque, Y. Kobayashi, Y. Kuno, Attention control system considering the target person's attention level, in: Proceedings of the 8th ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'13, IEEE Press, Piscataway, NJ, USA, 2013, pp. 111–112. [Online]. Available: http://dl.acm.org/citation.cfm?id=2447556.2447592.

[163] M.M. Hoque, T. Onuki, D. Das, Y. Kobayashi, Y. Kuno, Attracting and controlling human attention through robot's behaviors suited to the situation, in: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'12, ACM, New York, NY, USA, 2012, pp. 149–150. [Online]. Available: http://doi.acm.org/10.1145/2157689.2157729.

[164] T. Onuki, T. Ishinoda, Y. Kobayashi, Y. Kuno, Design of robot eyes suitable for gaze communication, in: Proceedings of the 8th ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'13, IEEE Press, Piscataway, NJ, USA, 2013, pp. 203–204. [Online]. Available: http://dl.acm.org/citation.cfm?id=2447556.2447638.

[165] A. Moon, D.M. Troniak, B. Gleeson, M.K. Pan, M. Zheng, B.A. Blumer, K. MacLean, E.A. Croft, Meet me where i'm gazing: how shared attention gaze affects human–robot handover timing, in: Proceedings of the 2014 ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'14, ACM, New York, NY, USA, 2014, pp. 334–341. [Online]. Available: http://doi.acm.org/10.1145/2559636.2559656.

[166] H. Admoni, A. Dragan, S.S. Srinivasa, B. Scassellati, Deliberate delays during robot-to-human handovers improve compliance with gaze communication, in: Proceedings of the 2014 ACM/IEEE International Conference on Human–Robot Interaction, ser. HRI'14, ACM, New York, NY, USA, 2014, pp. 49–56. [Online]. Available: http://doi.acm.org/10.1145/2559636.2559682.

[167] J.E. Hanna, S.E. Brennan, Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation, J. Memory Lang. 57 (4) (2007) 596–615.

[168] J.E. Hanna, M.K. Tanenhaus, Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements, Cogn. Sci. 28 (1) (2004) 105–115.

[169] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, N. Hagita, Footing in human–robot conversations: how robots might shape participant roles using gaze cues, in: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, ACM, 2009, pp. 61–68.

[170] L.B. Adamson, R. Bakeman, The development of shared attention during infancy, Ann. Child Dev. 8 (1991) 1–41.

[171] B. Scassellati, Mechanisms of shared attention for a humanoid robot, in: Embodied Cognition and Action: Papers from the 1996 AAAI Fall Symposium, vol. 4, no. 9, 1996, p. 21.

[172] B. Scassellati, Imitation and mechanisms of joint attention: a developmental structure for building social skills on a humanoid robot, in: Computation for Metaphors, Analogy, and Agents, Springer, 1999, pp. 176–195.

[173] G.O. Deák, I. Fasel, J. Movellan, The emergence of shared attention: using robots to test developmental theories, in: Proceedings 1st International Workshop on Epigenetic Robotics: Lund University Cognitive Studies, vol. 85, 2001, pp. 95–104.

[174] I. Fasel, G.O. Deák, J. Triesch, J. Movellan, Combining embodied models and empirical research for understanding the development of shared attention, in: The 2nd International Conference on Development and Learning, Proceedings, IEEE, 2002, pp. 21–27.

[175] M.W. Hoffman, D.B. Grimes, A.P. Shon, R.P. Rao, A probabilistic model of gaze imitation and shared attention, Neural Netw. 19 (3) (2006) 299–310.

[176] C. Peters, S. Asteriadis, K. Karpouzis, E. de Sevin, Towards a real-time gaze-based shared attention for a virtual agent, in: International Conference on Multimodal Interfaces, 2008.

[177] C. Peters, S. Asteriadis, K. Karpouzis, Investigating shared attention with a virtual agent using a gaze-based interface, J. Multimodal User Interfaces 3 (1–2) (2010) 119–130.

[178] D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind? Behav. Brain Sci. 1 (04) (1978) 515–526.

[179] H.M. Wellman, The child's theory of mind, 2011.

[180] B.M. Scassellati, Foundations for a theory of mind for a humanoid robot (Ph.D. dissertation), Massachusetts Institute of Technology, 2001.

[181] B. Scassellati, Theory of mind for a humanoid robot, Auton. Robots 12 (1) (2002) 13–24.

[182] L.F. Marin-Urias, E.A. Sisbot, A.K. Pandey, R. Tadakuma, R. Alami, Towards shared attention through geometric reasoning for human robot interaction, in: 9th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2009, IEEE, 2009, pp. 331–336.

[183] K. Santner, G. Fritz, L. Paletta, H. Mayer, Visual recovery of saliency maps from human attention in 3D environments, in: 2013 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2013, pp. 4297–4303.

[184] S. Baron-Cohen, Mindblindness: An Essay on Autism and Theory of Mind, MIT Press, 1997.

[185] S.E. Baron-Cohen, H.E. Tager-Flusberg, D.J. Cohen, Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience, Oxford University Press, 2000.

[186] B. Robins, K. Dautenhahn, R. Te Boekhorst, A. Billard, Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? Univers. Access Inform. Soc. 4 (2) (2005) 105–120.

[187] G. Bird, J. Leighton, C. Press, C. Heyes, Intact automatic imitation of human and robot actions in autism spectrum disorders, Proc. R. Soc. Biol. Sci. Ser. B 274 (1628) (2007) 3027–3031.

[188] B. Robins, P. Dickerson, P. Stribling, K. Dautenhahn, Robot-mediated joint attention in children with autism: a case study in robot–human interaction, Interact. Stud. 5 (2) (2004) 161–198.

[189] B. Robins, K. Dautenhahn, P. Dickerson, From isolation to communication: a case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot, in: Second International Conferences on Advances in Computer–Human Interactions, ACHI'09, IEEE, 2009, pp. 205–211.

[190] H. Admoni, B. Scassellati, Demo: toward a data-driven generative behavior model for human–robot interaction, in: Proceedings of the 2014 Workshop on Mobile Augmented Reality and Robotic Technology-Based Systems, ACM, 2014, pp. 19–20.

[191] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, PA, 2009.

[192] D. Jurafsky, H. James, Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics, and Speech, 2000.

[193] P.R. Cohen, C.R. Perrault, Elements of a plan-based theory of speech acts, Cogn. Sci. 3 (3) (1979) 177–212.

[194] J.J. Kuffner Jr., S.M. LaValle, RRT-connect: an efficient approach to single-query path planning, in: IEEE International Conference on Robotics and Automation, Proceedings. ICRA'00, vol. 2, IEEE, 2000, pp. 995–1001.

[195] S. Garrido, L. Moreno, M. Abderrahim, F. Martin, Path planning for mobile robot navigation using Voronoi diagram and fast marching, in: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2006, pp. 2376–2381.

[196] L.P. Kaelbling, M.L. Littman, A.R. Cassandra, Planning and acting in partially observable stochastic domains, Artif. Intell. 101 (1) (1998) 99–134.

[197] S. Young, Using POMDPs for dialog management. in: SLT, 2006, pp. 8–13.

[198] F. Bacchus, Lp, a logic for representing and reasoning with statistical knowledge, Comput. Intelligence 6 (4) (1990) 209–231.

[199] N. Mavridis, H. Dong, To ask or to sense? Planning to integrate speech and sensorimotor acts, in: 2012 4th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops, ICUMT, IEEE, 2012, pp. 227–233.

[200] V. Klingspor, J. Demiris, M. Kaiser, Human–robot communication and machine learning, Appl. Artif. Intell. 11 (7) (1997) 719–746.

[201] C. Breazeal, Imitation as social exchange between humans and robots, in: Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts, 1999, pp. 96–104.

[202] C.L. Isbell Jr., M. Kearns, S. Singh, C.R. Shelton, P. Stone, D. Kormann, Cobot in LambdaMOO: an adaptive social statistics agent, Auton. Agents Multi-Agent Syst. 13 (3) (2006) 327–354.

[203] E. Guizzo, Robots with their heads in the clouds, IEEE Spectr. 48 (3) (2011) 16–18.

[204] N. Mavridis, T. Bourlai, D. Ognibene, The human–robot cloud: situated collective intelligence on demand, in: 2012 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, CYBER, IEEE, 2012, pp. 360–365.

[205] N. Mitsunaga, Z. Miyashita, K. Shinozawa, T. Miyashita, H. Ishiguro, N. Hagita, What makes people accept a robot in a social environment-discussion from six-week study in an office, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2008, IEEE, 2008, pp. 3336–3343.

[206] N. Mavridis, A. AlDhaheri, L. AlDhaheri, M. Khanii, N. AlDarmaki, Transforming ibnsina into an advanced multilingual interactive android robot, in: 2011 IEEE GCC Conference and Exhibition, GCC, IEEE, 2011, pp. 120–123.

[207] M. Waibel, M. Beetz, J. Civera, R. D'Andrea, J. Elfring, D. Galvez-Lopez, K. Haussermann, R. Janssen, J. Montiel, A. Perzylo, et al., Roboearth, IEEE Robot. Autom. Mag. 18 (2) (2011) 69–82.

[208] D. Hunziker, M. Gajamohan, M. Waibel, R. D'Andrea, Rapyuta: the RoboEarth cloud engine, in: Proc. IEEE Int. Conf. on Robotics and Automation, ICRA, Karlsruhe, Germany, 2013.

[209] H. Sacks, E.A. Schegloff, G. Jefferson, A simplest systematics for the organization of turn-taking for conversation, Language (1974) 696–735.

[210] E.A. Schegloff, Overlapping talk and the organization of turn-taking for conversation, Lang. Soc. 29 (1) (2000) 1–63.

[211] Y. Matsusaka, S. Fujie, T. Kobayashi, Modeling of conversational strategy for the robot participating in the group conversation, in: INTERSPEECH, vol. 1, 2001, pp. 2173–2176.

[212] C. Chao, A.L. Thomaz, Turn taking for human–robot interaction, in: AAAI Fall Symposium on Dialog with Robots, 2010, pp. 132–134.

[213] C.L. Sidner, C. Lee, C.D. Kidd, N. Lesh, C. Rich, Explorations in engagement for humans and robots, Artif. Intell. 166 (1) (2005) 140–164.

[214] J.-M. Valin, F. Michaud, J. Rouat, D. Létourneau, Robust sound source localization using a microphone array on a mobile robot, in: 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2003, Proceedings, vol. 2, IEEE, 2003, pp. 1228–1233.

[215] K. Nakadai, D. Matsuura, H.G. Okuno, H. Kitano, Applying scattering theory to robot audition system: robust sound source localization and extraction, in: 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2003, Proceedings, vol. 2, IEEE, 2003, pp. 1147–1152.

[216] J.-M. Valin, F. Michaud, B. Hadjou, J. Rouat, Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach, in: 2004 IEEE International Conference on Robotics and Automation, Proceedings. ICRA'04, vol. 1, IEEE, 2004, pp. 1033–1038.

[217] J.-M. Valin, F. Michaud, J. Rouat, Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering, Robot. Auton. Syst. 55 (3) (2007) 216–228.

[218] K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, H. Tsujino, Design and implementation of robot audition system 'HARK'—open source software for listening to three simultaneous speakers, Adv. Robot. 24 (5–6) (2010) 739–761.

[219] D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. Speech Audio Process. 3 (1) (1995) 72–83.

[220] D.A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker verification using adapted Gaussian mixture models, Digit. Signal Process. 10 (1) (2000) 19–41.

[221] M. Ji, S. Kim, H. Kim, H.-S. Yoon, Text-independent speaker identification using soft channel selection in home robot environments, IEEE Trans. Consum. Electron. 54 (1) (2008) 140–144.

[222] Y. Matsusaka, T. Tojo, S. Kubota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, T. Kobayashi, Multi-person conversation via multi-modal interface-a robot who communicate with multi-user-. in: EUROSPEECH, vol. 99, 1999, pp. 1723–1726.

[223] K. Nakadai, D. Matsuura, H.G. Okuno, H. Tsujino, Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots, Speech Commun. 44 (1) (2004) 97–112.

[224] M. Katzenmaier, R. Stiefelhagen, T. Schultz, Identifying the addressee in human–human–robot interactions based on head pose and speech, in: Proceedings of the 6th International Conference on Multimodal Interfaces, ACM, 2004, pp. 144–151.

[225] H. Holzapfel, Towards development of multilingual spoken dialogue systems, in: Proceedings of the 2nd Language and Technology Conference, 2005.

[226] C. Cullen, C. Goodman, P. McGloin, A. Deegan, E. McCarthy, Reusable, interactive, multilingual online avatars, in: Conference for Visual Media Production, CVMP'09, IEEE, 2009, pp. 152–158.

[227] K.R. Echavarria, M. Genereux, D.B. Arnold, A.M. Day, J.R. Glauert, Multilingual virtual city guides, in: Proceedings Graphicon, Novosibirsk, Russia, 2005.

[228] T. Starner, J. Weaver, A. Pentland, Real-time American sign language recognition using desk and wearable computer based video, IEEE Trans. Pattern Anal. Mach. Intell. 20 (12) (1998) 1371–1375.

[229] C. Vogler, D. Metaxas, Handshapes and movements: multiple-channel American sign language recognition, in: Gesture-Based Communication in Human–Computer Interaction, Springer, 2004, pp. 247–258.

[230] G. Murthy, R. Jadon, A review of vision based hand gestures recognition, Int. J. Inf. Technol. Knowl. Manag. 2 (2) (2009) 405–410.

[231] H. Brashear, T. Starner, P. Lukowicz, H. Junker, Using multiple sensors for mobile sign language recognition, 2003.

[232] R. Plamondon, S.N. Srihari, Online and off-line handwriting recognition: a comprehensive survey, IEEE Trans. Pattern Anal. Mach. Intell. 22 (1) (2000) 63–84.

[233] T. Plötz, G.A. Fink, Markov models for offline handwriting recognition: a survey, Int. J. Doc. Anal. Recogn. (IJDAR) 12 (4) (2009) 269–298.

[234] L.M. Lorigo, V. Govindaraju, Offline arabic handwriting recognition: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 28 (5) (2006) 712–724.

[235] N. Mavridis, Autonomy, isolation, and collective intelligence, J. Artif. Gen. Intell. 3 (1) (2011) 1–9.

Nikolaos Mavridis received his Ph.D. from MIT in 2007, after receiving his M.S.E.E. from UCLA and M.Eng. in ECE from the Aristotle University of Thessaloniki. From 2007 to 2011, he served as Asst. Professor at the College of IT of the United Arab Emirates University, where he founded the Interactive Robots and Media Laboratory (IRML), and at New York University AD, and also affiliated with the NYU-Poly as Assistant Professor of Research, and is now at NCSR Demokritos in Athens. His lab is home to the "FaceBots" social robots project, as well as to "IbnSina", the first arabic-language conversational android robot. In his Ph.D. Thesis at MIT, he introduced the "Grounded Situation Model" proposal, and demonstrated its benefits by implementing it on Ripley, a manipulator robot with vision, touch, and speech synthesis. The sensory motor/linguistic abilities of the resulting system were comparable to those implied by a standard psychological test for 3-year old children (The "Token Test"). The research interests of Dr. Mavridis include Robotics, especially Interactive and Social Robotics, as well as Cognitive Systems.