In J. Markowitz (Ed.), Robots That Talk and Listen. Boston, MA: De Gruyter.

From Talking and Listening Robots to Intelligent Communicative Machines

Roger K. Moore
University of Sheffield

Abstract

It is a popular view that the future will be inhabited by intelligent talking and listening robots with whom we shall converse using the full palette of linguistic expression available to us as human beings. Of course, recent technical and engineering developments such as Siri would appear to suggest that important steps are being made in that direction – and indeed they are. However, it is argued here that we need to go far beyond our current capabilities and understanding towards a more integrated perspective; simply interfacing state-of-the-art speech technology with a state-of-the-art robot is very unlikely to lead to effective human-robot interaction. We need to move from developing robots that simply talk and listen to evolving intelligent communicative machines that are capable of truly understanding human behavior, and this means that we need to look beyond speech, beyond words, beyond meaning, beyond communication, beyond dialog and beyond one-off interactions.

I. Introduction

The idea that humanity's future world will be populated by intelligent robots, and that we will converse with them in exactly the same way that we interact with human beings, is a compelling image in contemporary science fiction. From *Star Wars'* golden android C-3PO to Pixar's rusty dust-busting WALL•E, we have a collective vision that it is only a matter of time before such man-made artifacts are an everyday reality. If only this were true. If scientists and engineers really knew how to make such communicative devices, then not only would we have realized the vision of many futurist thinkers (Gates Myhrvold and Rinearson 1995, Kurzweil 1999 and Kurzweil 1990), but we will have also made deep inroads into our understanding of what it means to be a human being.

The reason this is the case is that, contrary to that which is commonly portrayed in the popular media (for example, in *Her* – the 2013 American science-fiction, romantic-comedy film), creating an autonomous agent that is capable of interacting effectively with human beings using the full palette of linguistic expression requires us to go far beyond our current understanding and capabilities in robotic systems and spoken-language processing. It turns out that science fiction is just that – *fiction*! In our enthusiasm to reach some kind of utopian future populated by conversational devices, we have overlooked one of the deepest puzzles faced by humankind – how (and *why*) do we communicate with each other in the first place (a much studied, but still open question - Fitch 2000, Hockett 1960, Holden 2004, MacNeilage 2008, Tomasello 2008), and how can we simulate such behavior in computer-based machines that need to be able to function in real-world environments (Feldman 2008 and Moore 2007c)?

Not only is progress hampered by our lack of understanding, but it is also impeded by the fact that many roboticists regard a speech-enabled interface as a somewhat independent, bolt-on goody rather than a natural extension of a robot's perceptuo-motor system (Schwartz et al. 2012). The assumption seems to be that speech-technology components should be available off-the-shelf in order to provide an instant, voice-based interface for almost any robot platform. Indeed, the very word "interface" implies a degree of peripheral independence between the so-called speech engines and the remainder of the system. This situation is all the more surprising given that it is now generally recognized in contemporary 'enactive' robotic circles (Vernon 2010) that embodiment, behavioral grounding and situated awareness appear to be key to resolving perceptual uncertainty and motor behavior planning; and that these issues can be addressed by a commitment to multimodal, sensorimotor integration and cognitively-motivated processing (Vernon, Metta and Sandini 2010). However, notwithstanding a few exceptions (Moore 2007a and Pickering and Garrod 2013), it is rather rare for these themes to be applied to the spoken-language channel.

Of course, recent commercial developments in spoken-language recognition and understanding (such as Siri - Apple's speech-operated, personal assistant and knowledge navigator, released in 2011) represent immense technical and engineering achievements. Indeed, such developments come less than 15 years after the first commercial release of *continuous,*[1] automatic speech-recognition systems (Pieraccini 2012). However, by opening up the opportunity to use the voice channel, system designers are inadvertently entering a parallel world where a user's natural behaviors can soon undermine any productive outcomes. Interaction based on speech and language appears to be all-or-nothing; it seems to be very difficult for a naïve user to accommodate the type of prescribed sub-language (as is used, for example, by the military, air-traffic control or in some of the professions) that current spoken-language technology depends upon. Indeed, there is now a rather hollow ring to one manufacturer's advertising slogan from the 1990s – "*You have been learning since birth the only skill needed to operate our equipment*." It turns out that speech-based interfaces are anything but natural, and there appears to be a fundamental mismatch between the capabilities of contemporary speech-enabled devices and the intuitive behaviors of their potential users

[1] Speech recognizers that can accommodate running speech rather than words isolated by artificial pauses.

(as evidenced by the large number of humorous videos of miscommunication with Siri that have been posted on channels, such as YouTube and in the press.[2] There is even a dedicated website[3]).

On the other hand, it has to be said that, whilst human-human, speech-based interaction is incredibly effective (even in extremely noisy or challenging conditions); it is by no means perfect. People often misrecognise or misunderstand each other, or have difficulty with unfamiliar voices or accents, or fail to speak clearly or express themselves properly, or mispronounce important information. As Larry Rabiner (a retired senior speech researcher at AT&T Bell Labs.) remarked at the 1997 IEEE International Workshop on Automatic Speech Recognition and Understanding: "*If speech communication between people is so effective, how come are there so many lawyers?*" (Rabiner 1997).

The gap between human-human, spoken-language interaction and speech-based, human-machine dialog is thus not one that is adequately characterized by word accuracy or task-completion times. It seems that the former is a highly evolved channel of general-purpose, communicative interaction between individuals that is extremely tolerant to the demands and quirks of everyday reality, whereas the latter is a rather restrictive medium that can easily degrade catastrophically when there is even the smallest deviation from expected behaviors. It is this difference that needs to be addressed if we are to make substantive progress towards designing and implementing talking and listening robots which seek to engage productively (and intuitively) with their human interlocutors.

This Chapter addresses these issues by decomposing the solution space along quite different lines to approaches that have been attempted hitherto. It tries to place speech in a broader theoretical framework that militates against it being seen as an independent faculty that can simply be grafted onto a pre-existing 'intelligent' agent (such as a robot) as an optional extra or nice-to-have technical feature. Rather, it is argued that speech-based interaction is fundamental to the behavior of healthy human beings, and we ignore – at our peril – the role that spoken language plays in shaping our world along with our actions and interactions within it. We need to move from developing robots that simply talk and listen, to evolving intelligent, interactive machines that communicate, understand and act in real-world contexts that are co-inhabited by other living, communicative systems such as human beings, animals and pets.

II. Looking for Solutions

Whilst it is very easy to criticize contemporary approaches to voice-enabled robots for taking a superficial stance with respect to the complexities of speech and language, it has to be acknowledged that our current limited state of knowledge inevitably pushes research towards the more pragmatic and

[2] For example, http://bcove.me/1m4r05ow

[3] http://www.sirifunny.com/

obvious solutions. Part of the problem is that speech and language straddles so many disparate academic disciplines, ranging from psychology, phonetics and linguistics, to computer science, engineering and human-computer interaction. Very few individuals have formal training in all of the relevant fields of study, and this has led to a remarkable diversification of perspectives on what is essentially a single phenomenon.

These issues come to the fore when we consider the role of spoken language in the context of human-robot interaction. Robots, by their very nature, are immensely complex physical and electrical machines; even arranging for one to move in a coordinated and purposeful manner poses enormous engineering challenges (Arkin 1998 and Winfield 2012). Sensor systems are often limited, and yet present a complex array of high-dimensional, potentially noisy signals to the underlying control architectures. Algorithms for interpreting sensor data, and for planning and executing actions, are still the subject of much research. Constructing a robot to navigate and interact in a static and controlled environment is a major undertaking; developing one to operate within a dynamic, real-world environment inhabited by living organisms with their own aims, objectives and established communication channels is an enormous challenge. In this context, it is not surprising that the temptation to integrate an off-the-shelf automatic speech recognizer and/or text-to-speech synthesizer is simply too great to resist.

In order to mitigate some of these problems, it is necessary to understand the context in which spoken language normally operates, together with the role it plays in facilitating interaction between intelligent agents, particularly human beings. It is time to view spoken language as

- *not* a standalone faculty that is independent of other sensorimotor channels;
- *not* a peripheral behavior that is independent of core cognitive processes;
- *not* an activity that is independent of the real-world context in which it takes place;
- *not* a one-off exchange with no prior history.

If we are to make progress, it seems that we need to look beyond these traditional perspectives.

A. Beyond speech

Whilst it is clearly possible for one person to communicate with another reasonably well via a purely acoustic channel (for example, when using the telephone), this is obviously not the normal configuration. Spoken language has evolved as part of a multimodal complex of interactive behaviors involving features such as overall appearance, body posture, facial expressions, eye gaze, gestures and pointing (Esposito and Esposito 2011 and Mehrabian 1968). It appears that communicative information is not only distributed across these channels in a coordinated and coherent manner, but it is also optimized according to the physical and temporal context in which an interaction takes place. For example, just as a person (or an animal) will walk around a physical object that impedes its intended path, so words are chosen to clarify potentially obscure communicative points; the speed, loudness and clarity of speech are all adapted to the characteristics of the environment in which it is produced (Lindblom 1990 and Lombard 1911); and the delivery of linguistic expressions is coordinated with physical movements and gestures (Wagner, Malisz and Kopp 2014). Even choosing the appropriate moment to speak is an actively-managed, context-dependent behavior (Clark 2002).

Of course this does not mean that information may be arbitrarily distributed between speech and other communicative channels. Spoken language confers a number of adaptive benefits to the human organism: it represents a remarkably high data-rate link (of the order of 100 bits-per-second) compared to other communication channels; it can operate over considerable distances or in the dark; it allows communication to take place even if the hands or eyes are engaged with other tasks; it provides a formal mechanism for expressing and manipulating high-level conceptual representations; and it facilitates the abstraction (and hence, generalization) of thoughts and ideas.

The implication of all this is that human speech is a cognitively-based faculty that is fully integrated and finely synchronized with other sensorimotor activities and that, like other behaviors, it is actively and dynamically managed to meet the needs of the communicative context. Overall, the behavior of healthy, living systems is mostly coordinated and coherent, and any deviation from such consistency may be interpreted as physical or mental illness. Unfortunately, this is exactly the situation that applies to many robotic systems: mismatched capabilities in terms of what a robot looks like, what it sounds like, what it says or how it behaves are a recipe for, at best, confusion on the part of a user or, at worst, *repulsion*. Indeed, it has been shown by the author that the 'uncanny valley effect' (Mori 1970) may be explained by conflicting cues giving rise to perceptual tension at category boundaries (Moore 2102), and the importance of achieving consistency is supported by a wealth of empirical evidence (for example, Bartneck et al. 2009, Komatsu and Yamada 2007, Otsuka et al. 2010 and Walters et al. 2008) including matching faces and voices (Mitchell et al. 2011 and Moore and Maier 2012).

What this means is that designers of talking and listening robots need to approach human-robot interaction from an integrated and coordinated, multimodal perspective. A robot's look, sound and behavior need to be coherent and representative of the actual capabilities of the device. The distribution of communicative information needs to be actively managed across the different modes taking into account the characteristics of the environmental context; and all behaviors need to adapt to the communicative context as part of a dynamic optimization process.

B. Beyond words

Notwithstanding the 2011 release of Apple's Siri, there has been an understandable tendency within the speech-technology, R&D community to focus heavily on the recognition and synthesis of *words*. This is not so surprising since words represent a reasonably well-defined ground truth for the surface content of a spoken utterance. This has enabled the field to progress through the establishment of quantitative-performance benchmarks and competitive challenges (Pieraccini 2012). However, it is widely acknowledged that, aside from purely transcription-based applications (such as machine-based voice dictation); the real communicative target is not words but *meanings*.

In fact, in some senses the automatic recognition of spoken words is a solved problem, mainly thanks to the introduction of dynamic-programming search and hidden Markov modeling in the 1970s and 80s (Gales and Young 2007). Yes, it's unreliable in realistic environments; yes, it fails badly for regional accents; yes, it makes more mistakes if you speak too clearly[4]; but these are engineering challenges with no shortage of potential solutions (see the proceedings of any INTERSPEECH[5] or ICASSP[6] conference). What is far more difficult is going beyond just trying to determine what words are being spoken to figure out *why* someone has said what you think they might have said and what you're supposed to do about it.

Likewise, it is relatively straightforward to configure a speech synthesizer to read out a defined sentence with a selected voice-type and prescribed prosodic contour. What is more challenging is determine what to say, when to say it, and how such choices are manifest in the way in which an utterance is to be spoken. However, some interesting steps are being taken in this direction. For example, the recent introduction of hidden Markov model-based speech generation (Zen, Tokuda and Black 2009) has paved the way for a novel form of *reactive* speech synthesizer that is able to adapt its behavior as a function of the listener's apparent understanding – initially speaking more clearly in noise (Moore and Nicolao 2011), but ultimately selecting what to say (and how to say it) in order to achieve its communicative goals (Moore 2007b).

Nevertheless, understanding is much more than determining what to say next (and how to say it). The representation of meaning is a traditional (and challenging) area of research in the field of Artificial Intelligence, and there is not space here to review the vast literature on the subject. Suffice to say that the implications for spoken language have been appreciated for over 40 years (Newell et al. 1973). Yet no great inroads had been made into this area until the appearance of Siri revealed the power of online access to 'big data' for facilitating the interpretation of general, spoken enquiries. Siri represents an important

[4] This well-known phenomenon is a direct consequence of treating speech as a pattern to be recognised rather than as a communicative signal to be *understood*.

[5] http://www.isca-speech.org/iscaweb/index.php/conferences/interspeech

[6] http://ieeexplore.ieee.org/xpl/conhome.jsp?punumber=1000002

step towards demonstrating the practical value of going beyond the words of an utterance, but it is only a small step towards unraveling the true potential of spoken interaction.

Of particular significance to the use of spoken language to communicate with robots is the potentially important fact that, unlike personal agents and avatars (such as Siri), robots are instantiated as physical entities in the real world. This means that a robot's behaviors are necessarily grounded in (and constrained by) the characteristics of its environment. This has led to a line of thought which claims that *real* meaning is grounded in bodily experience rather than in a prescribed ontology of logical forms (Jirak et al. 2010). Of course, spoken language (unlike written language) is also grounded in the real-world, and this perspective has led to the hypothesis that the meaning of language is also based on bodily experience and that understanding is mediated by the use of metaphor as a mechanism for generalization (Feldman 2008 and Lakoff and Johnson 1980). These ideas intersect nicely with recent developments in 'action understanding' which suggest that interpreting the behaviors and emotions of others is not based on a process of logical reasoning (as typified by traditional GOFAI[7] approaches), but by direct simulation of the observed events (Gallese, Keysers and Rizzolatti 2004). Indeed, these ideas are supported by neurological evidence, driven largely by the discovery of so-called 'mirror neurons' (Kohler et al. 2002 and Rizzolatti and Craighero 2004), and give rise to the key claim that the core meaning of a word, such as "grasp," is the complex neural circuitry that supports the actual action of grasping (Rizzolatti and Arbib 1998). Such a perspective suggests that, by virtue of operating in the real world, physically-embodied robots would be in a much more powerful position to understand and manipulate language than disembodied avatars or personal agents (Billard and Dautenhahn 1998 and Marocco et al. 2010).

These ideas have been picked up by the mainstream speech community, and have led to the development of the PRESENCE (PREdictive SENsorimotor Control and Emulation) architecture (Moore 2007b) and PACT (Perception-for-Action-Control Theory; Schwartz et al. 2012). Both of these approaches speculate on the potential value of close links between speech perception and speech production. The PRESENCE model introduces the notion of speech recognition-by-synthesis as well as speech synthesis-by-recognition.

C. Beyond meaning

One might be forgiven for thinking that once the meaning of a spoken utterance has been determined, all a robot has to do is to act upon its contents and respond appropriately. In reality, this is just the beginning of the story. Living systems such as human beings are complex cognitive agents whose behaviors are determined by their individual drives and needs as well as by their beliefs and intentions. If

[7] 'Good Old-Fashioned Artificial Intelligence' (Haugeland 1985)

a robot is to interact successfully with a human being, then it can only do so effectively by modeling such conditioning variables, either in general or, more usefully, on a user-specific basis. User modeling is a familiar concept in human-computer interaction (Benyon 2010, Dix et al. 2004 and Rogers, Sharp and Preece 2011), but such models typically retain user preferences rather than acknowledge the deeper motivational factors that drive a user's behavior and shape their verbal utterances.

For a talking and listening robot these issues bear directly on the importance of being able to interpret and express the *paralinguistic* phenomena which arise as an emergent property of interaction between two or more teleogically-based entities. In the condition where two agents (such as a human and a robot) have convergent aims and objectives, the effort required to perform a joint task may be shared between the participants. This situation facilitates the expression of satisfaction and pleasure by both agents (for example, by smiling or emitting appropriate sounds) – behaviors which can also color spoken language in both form and content and which serve to support further social cooperation.

On the other hand, in the condition where two agents have divergent aims and objectives (perhaps as a result of previous misunderstandings), then the efforts required to perform a joint task may be vastly increased, potentially leading to conflict between the participants which, in turn, may be realized as displays of displeasure signaling the failure to cooperate effectively (and coloring behaviors such as speech accordingly).

What this tells us is that, in order for a robot to derive a clear understanding of what is meant by any particular utterance, it is necessary not only to determine the words that are being said, but also the way in which they are being said. Spoken language is far more subtle in its communicative function than is implied by its literal interpretation. If a robot is to truly engage with a human being using language, then it must have at least a primitive capacity to perceive and express appropriate affective cues. Luckily, the interpretation and expression of emotion has been a hot topic of research since the publication of Picard's book on 'Affective Computing' (Picard 1997), and there has since been considerable interest in the consequences for speech and spoken language interaction (e.g. André et al. 2004, Brück, Kreifelts and Wildgruber 2011, Scherer 2003, Simon-Thomas et al. 2009 and Vogt, Andre and Bee 2008), multimodal interaction (Esposito 2009) and for affective human-robot interaction (Breazeal 2000, Breazeal and Aryananda 2002 and Kirby, Forlizzi and Simmons 2010).

D. Beyond communication

In recent years, research into affective science has been extended to include more general aspects of social signaling (Pentland 2008 and Vinciarelli, Pantic and Bourlard 2009) and social engagement (Payr et al. 2009). The overall objective is seen as one of bridging the social-intelligence gap between humans and machines (Vinciarelli et al. 2012) by establishing design principles for creating social agents (Gorbunov, Barakova and Rauterberg 2013). The realization is that interaction between living organisms in general, and social species (such as human beings) in particular, is actually much richer than word signaling (the coding and decoding of messages) implies. Rather, an individual's behaviors are crafted to support continuous, coordinated interactions within or between social groupings. Information is not just passed from one individual to another; available communication channels are exploited to manipulate the behavior of others for cooperative or competitive ends. Likewise, information is not packaged into discrete communicative nuggets; rather behaviors are aligned to provide continuous, adaptive coupling between individuals – leading to an emergent phenomenon that Stephen Cowley describes as 'co-action' (Cowley and Macdorman 2006: 363).

All of this is neutral with respect to the modes and channels exploited by living systems to achieve their desired goals. Whether such coordinated behavior is achieved by means of visual or vocal display is not the prime issue. However, the facts that the vocal channel offers certain advantages in particular situations and environments, and that spoken language provides an exceptionally high information rate (both of which were discussed earlier), mean that it is not surprising that the dynamics of coordination between individual human beings relies heavily on spoken language (Cowley 2009). Hence, for successful human-robot interaction, artificial systems need to be able to model the co-active coupling between talking listeners and listening talkers in concert with all other modes of interaction, including the shifting of attention and focus from one mode to another.

E. Beyond dialog

Some progress is being made in these areas. The traditional notion of strict turn-taking between a user and a spoken language dialog system is giving way to a more fluid interaction based on partial hypotheses and *incremental* processing (Hastie, Lemon and Dethlefs 2012, Skantze and Schlangen 2009). Likewise, simple slot-filling approaches to language understanding and generation are being replaced by sophisticated statistical methods for estimating dialog states and optimal next moves (Gasic et al. 2013, Williams and Young 2007).

However, interaction between higher living organisms such as human beings is more than dynamic coupling and more than step-by-step negotiation. Interaction, especially in social groups, is grounded in the relationships that exist (or are developed) between individuals. Interaction depends on the social status of the participants, the dominance relations between one individual and another and the trust that individuals put in each other. These relations not only condition the dynamics of interaction, but they also act as priors on the selected behaviors themselves, i.e. they influence the way in which people talk. This means that a voice-enabled robot cannot be viewed as a neutral partner in a spoken-language dialog; rather its perceived social status and perceived believability will strongly influence the way in which users attempt to interact with it. People typically employ small-talk to establish such relations, so it is likely that such strategies would be important for effective human-robot interaction (Bickmore and Cassell 2001). Failure to establish such relations at an early stage in a dialog/interaction could lead to user confusion and the collapse of the interaction (Nass and Brave 2005).

Another significant factor that influences the trajectory of an interaction is whether or not an agent is perceived as intentional, i.e. an object with its own motives and goals, rather than one that simply follows the laws of physics or just does what it is designed to do (Dennett 1989). If a user takes a *design* stance to an object or device, then any unexpected behavior is taken to indicate that the device is broken and interaction should be abandoned. However, if a user takes an *intentional* stance to a device (which is highly likely for a robot), then any unexpected behavior is taken as evidence that there are hidden motivations and goals that need to be determined and perhaps changed. In the latter circumstances, users are effectively assigning a 'theory of mind' (Premack and Woodruff 1978) to the agent/robot, and interaction is unlikely to proceed if the robot is unable to explain its hidden mental states adequately (Brooks, Smith and Scassellati 2001). Indeed, the easiest way for a robot to be perceived as intentional, is for a robot to *be* intentional, i.e. to have its own internal needs and goals driving its behavior.

F. Beyond One-off Interactions

Finally, a critical issue that strongly conditions spoken-language behavior is that participants usually have a considerable prior history of interaction. People are familiar with the voices of their family, friends and colleagues. They retain considerable person/context-specific memories of previous conversations, and are able to draw on this information in order to interact efficiently[8]. In contrast, most human-robot interactions are short-term, with little or no memory (in the robot) of previous encounters. Clearly speech-based applications such as Siri are moving towards personalized interactions with strong user models and some retained history, and there is a realization that robots need to have a similar long-term perspective (Gockley et al. 2005, Payr et al. 2009, Wilkes 2010 and EU STRANDS project[9]). Bringing these two areas together would seem to be an important step towards more effective speech-based human-robot interaction and towards a more cognitively-motivated user interface.

[8] There is an apocryphal story of elderly cousins meeting after several years apart living on different sides of the world. The long overdue conversation started with the utterance "*As I was saying* …"!

[9] STRANDS stands for 'Spatio-Temporal Representation and Activities for Cognitive Control in Long-Term Scenarios': see http://strands.acin.tuwien.ac.at/index.html

> *"… future generations of computer-based systems will need cognitive user interfaces to achieve sufficiently robust and intelligent human interaction … characterized by the ability to support inference and reasoning, planning under uncertainty, short-term adaptation, and long-term learning from experience."* (Young 2010: 128).

The benefit of providing a talking and listening robot with a long-term memory of previous encounters is not only the facilitation of personalized conversational interaction; but it is also the opportunity to consolidate information from memory for the purpose of generalizing to novel situations with known users or generalizing to novel users in known situations. Indeed, there is a community of researchers who believe that a robot may only be able to handle sophisticated, language-based interaction if it has acquired the requisite skill through the accumulated memory of its own personalized experience (Dautenhahn 2004, Demiris and Meltzoff 2008 and Schmidhuber 2006). In other words, it is hypothesized that a robot may have to learn spoken language by following the same developmental trajectory as a child (Dominey and Boucher 2005, Cangelosi et al. 2010, Serkhane, Schwartz and Bessière 2005 and ten Bosch et al. 2009).

III. Towards Intelligent Communicative Machines

The discussion thus far has highlighted a number of key areas where it seems that we need to extend the state-of-the-art if we are to create effective talking and listening robots. Pointers have been given to potential new directions in several different technical areas, and links have been established between various contributing disciplines. This section attempts to bring all these ideas together into a design framework that could take us beyond talking and listening robots towards intelligent, communicative machines.

Before launching into a consolidated perspective, however, there is one outstanding issue to be discussed. There is a natural tendency to assume that we need to produce a talking and listening robot that is effectively a facsimile of a fully-grown adult human being. In fact, there is considerable interest in being able to do that just in terms of a robot's overall look and behavior (e.g. Nishio, Ishiguro and Hagita 2007). Putting aside for one moment the risk of entering the uncanny valley (as discussed earlier), the reality is that achieving this level of ability across *all* aspects of a robot's design is currently impossible, and may well remain impossible for many years[10]. This means that we are faced with a real dilemma: to create a robot which gives the appearance of having the requisite human-level cognitive, linguistic and behavioral abilities but is, in reality, limited; or to design a robot whose appearance reflects its genuine, but limited, cognitive, linguistic and behavioral abilities.

[10] Notwithstanding the predictions of high-profile pundits such as Ray Kurzweil.

Interestingly, the vast majority of research in this area appears to favor the imitation approach, i.e. the creation of behaviors in a robot or artificial agent that mimic those observed in human beings in the hope that this will create a sufficient suspension-of-disbelief for interaction to proceed (Nass and Brave 2005). However, in the author's view this approach is fundamentally dishonest, and goes against the recommendations of the emerging Roboethics community (e.g. Boden et al. 2011). On the other hand, rather few researchers seem to advocate the second approach, i.e. the appropriate scaling of behaviors in a robot or artificial agent in order to create a coherent usability sub-space (Balentine 2007). The arguments laid out here suggest that only the latter approach can lead to sustainable progress; purely imitative designs, whilst entertaining in the short-term, have no long-term value and even serve to displace effort that could be directed towards real progress.

A. Achieving an Appropriate Balance of Capabilities

Of course, the reason that designing scaled-down capabilities is currently a minority interest is that, in practice, it is very hard to figure out how to achieve the necessary balances across the different cognitive, linguistic and behavioral dimensions. In effect, what is being asked for is a low-fidelity solution in which all the requisite capabilities are matched in accordance with some as yet unspecified (and ill-understood) metric. Perhaps suitable insights and guidance can be gained by considering the ways in which people interact successfully with their pets - even using speech - despite the large mismatch between a human being's abilities and those of, say, a dog (Lakatos et al. 2012).

Another place to look for insights into how to create a balanced set of interactive capabilities is in the entertainment media, especially animated cartoons (Moore 2011). Whilst it has been argued above that science fiction isn't necessarily the most reliable pointer to future technologies, it has to be said that the characteristics of talking and listening objects (including robots) in such media are often very carefully designed to elicit appropriate and believable characters. Indeed, the film *WALL•E* not only portrays voices that are clearly appropriate to the individual robots and machines involved, but it also presents a very important contrast between the awkward rusty metal-based male hero robot with his awkward rusty metal-based male-sounding voice and the slick smooth force-field-based female robot and her slick smooth force-field-based female-sounding voice[11]. Likewise, at a more cognitive level, the talking toaster in the UK TV science-fiction, comedy drama *Red Dwarf* is a brilliant characterization of a one-dimensional, bread-obsessed AI-based, electrical appliance whose entire perception of the world, and thus its conversation, revolves entirely around its need to provide "*hot buttered scrummy toast*" (Moore 2013).

[11] Ben Burtt Special: WALL•E – The Definitive Interview: http://designingsound.org/2009/09/ben-burtt-special-wall-e-the-definitive-interview/

B. A Consolidated Perspective

All these considerations may be summarized in a set of technical requirements that would seem to be essential if we are to move beyond simple talking and listening robots to more sophisticated intelligent communicative machines that are able to engage in effective speech-based interaction with real users in real environments.

- An intelligent communicative machine's skills should be dynamic and adaptable, i.e. they should be revised and updated constantly as the result of each and every interaction that takes place (e.g. by acquiring new words, new meanings, new ways of expressing concepts and new strategies for conversing).
- An intelligent communicative machine should retain a long-term record of all interactions (i.e. all sensorimotor data and associated internal states), thereby allowing it to draw on an extensive autobiographical memory for contextualizing each subsequent interaction.
- An intelligent communicative machine's behaviors should be driven by a prescribed set of internal needs and intentions which are directed towards servicing the needs and intentions of its potential users.
- All aspects (visual, vocal, cognitive and behavioral) of an intelligent communicative machine should be coherent as well as being both representative and informative of its actual abilities and its social status in a conversational context.
- Interaction with an intelligent communicative machine should be continuous, fluid and co-active such that there is no wrong time for a user to speak, listen or anything else.
- An intelligent communicative machine must have at least a primitive capacity to perceive and express appropriate affective cues.
- An intelligent communicative machine should seek to obtain a model of its user's beliefs, desires and intentions.
- An intelligent communicative machine should exploit its embodiment and provide a clear (centralized or distributed) physical presence to its users.
- An intelligent communicative machine should actively manage (using a dynamic optimization process) the distribution of information across the different sensorimotor channels as a function of the users' abilities and the environmental context.

C. Beyond human abilities

Finally, although it might seem premature given the current state-of-the-art in spoken-language technologies, it is tempting to speculate on the peculiar benefits that might be accrued by future intelligent communicative machines beyond those already exhibited by human beings. An obvious possibility is that such a machine might one day be able to recognize speech more accurately (and in more difficult environments) than a human being – a capability that speech researchers at IBM refer to as 'super-human speech recognition' (Picheny and Nahamoo 2008). However, although performing a task such as speech recognition better than a human sounds attractive, it is debatable whether such an ability is actually feasible given the human-centric nature of spoken language. What is much more interesting is the possibility of a machine doing something that a human being simply cannot do.

For example, a huge disadvantage of a living system is that all its parts are collocated; it exists in one place and its sensors and actuators are only able to function within a local area. A machine, on the other hand, has no such restrictions; its sensors and actuators may be distributed widely – throughout a home, across a city or around the planet. This means that an intelligent communicative machine's eyes, ears and mouths can literally be anywhere and everywhere. Obviously swarms of living systems in general or communities of human beings in particular, share some of the same properties, but these systems are limited by the properties of the communication channels that bind them together, and they contain no cognitive core capable of consolidating information across all channels. A machine would have none of these limitations; it could have longer long-term memory, it could share all information amongst all agents and it could hold multiple conversations at the same time. Sound familiar? These are the core characteristics of *Star Trek's* Borg Collective!

Is such a future desirable or undesirable? That is an appropriate question for society to debate. Is such a future technically feasible? Yes it is – so that debate should commence sooner than later.

IV. Conclusion

This chapter has attempted to bring together a number of different perspectives on talking and listening robots. The main thrust has been to demonstrate that simply adding off-the-shelf speech technology to an off-the-shelf robot is unlikely to create a usable or productive human-robot interface. Rather, it has been argued that we need to go beyond the state-of-the-art in a number of key areas: beyond speech, beyond words, beyond meaning, beyond communication, beyond dialog and beyond one-off interactions. It has been shown that it is better to design a robot with genuine but limited cognitive, linguistic and behavioral abilities than one which gives the appearance of having the requisite human-level abilities but, in fact, does not. It has also been suggested that the robots portrayed in popular films such as WALL•E may provide useful insights into how to create a balanced set of interactive capabilities. From these considerations a set of technical requirements are derived that would enable us to move from simple talking and listening robots to more sophisticated, intelligent communicative machines. Finally, it is argued that the challenge is not to emulate a human being (or to create a super-human being) but to extend capabilities in ways that are impossible for human beings.

Acknowledgements

I'd like to extend my thanks to my colleagues Dr. Peter Wallis for numerous discussions on many of the topics I have raised. Also, during the period I was preparing this chapter, I have received partial support by the European Commission under grant agreement number FP7-ICT-2013-10-EASEL – 'Expressive Agents for Symbiotic Education and Learning.'

References

André, E., Rehm, M., Minker, W. and Bühler, D. (2004) 'Endowing Spoken Language Dialogue Systems with Emotional Intelligence.' in *Proceedings Affective Dialogue Systems Conference*, LNAI 3068. held 14-16 June 2004 at Kloster Irsee, Germany. Berlin, Germany: Springer Science+Business Media, 178-187.

Arkin, R.C. (1998) *Behavior-Based Robotics*. Cambridge, MA USA: The MIT Press.

Balentine, B. (2007) *It's Better to Be a Good Machine than a Bad Person: Speech recognition and other exotic user interfaces at the twilight of the Jetsonian Age*. Annapolis, MD USA: ICMI Press.

Bartneck, C., Kanda, T., Mubin, O. and Al Mahmud, A. (2009) 'Does the Design of a Robot Influence Its Animacy and Perceived Intelligence?' *International Journal of Social Robotics* 1(2), 195–204.

Benyon, D. (2010) *Designing Interactive Systems*. 2nd edn. Boston, MA USA: Addison-Wesley.

Bickmore, T. and Cassell, J. (2001) 'Relational Agents: A model and implementation of building user trust.' in *Proceedings of ACM CHI 2001*. held 31 March - 5 April, 2001 at Seattle, OR USA, New York, NY USA: ACM, 396-403.

Billard, A. and Dautenhahn, K. (1998) 'Grounding Communication in Autonomous Robots: An experimental study.' in Recce, M. and Nehmzow. U. (eds.), *Robotics and Autonomous Systems, Special Issue on Scientific Methods in Mobile Robotics* 24, 71–79.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorell, T., Wallis, M., Whitby, B. and Winfield, A. (2011) *Principles of Robotics* [online] Swindon, UK: The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC). available from <http://www.epsrc.ac.uk/research/ourportfolio/themes/engineering/activities/Pages/principlesofrobotics.aspx> . [22 December 2013].

Breazeal, C. (2000) *Sociable Machines: Expressive social exchange between humans and robots*. Unpublished PhD thesis. Cambridge, MA USA: MIT [online] available from <http://groups.csail.mit.edu/lbr/mars/pubs/phd.pdf >. [10 January 2013].

Breazeal, C. and Aryananda, L. (2002) ‘Recognition of Affective Communicative Intent in Robot-Directed Speech.’ *Autonomous Robots* 12(1), 83-104.

Brooks, R., Smith, A.C. and Scassellati, B. (2001) *Foundations for a Theory of Mind for a Humanoid Robot*. Cambridge, MA USA: MIT.

Brück, C., Kreifelts, B. and Wildgruber, D. (2011) ‘Emotional Voices in Context: A neurobiological model of multimodal affective information processing.’ *Physics of Life Reviews* 8(4), 383–403.

Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., Tani, J., Belpaeme, T., Sandini, G., Nori, F., Fadiga, L., Wrede, B., Rohlfing, K., Tuci, E., Dautenhahn, K., Saunders, J. and Zeschel, A. (2010) ‘Integration of Action and Language Knowledge: A roadmap for developmental robotics.’ *IEEE Transactions on Autonomous Mental Development* 2(3), 167-195.

Clark, H.H. (2002) ‘Speaking in Time.’ *Speech Communication* 36(1-2), 5–13.

Cowley, S.J. (2009) ‘Distributed Language and Dynamics.’ *Pragmatics & Cognition* 17(3), 495–508.

Cowley, S.J. and Macdorman, K.F. (2006) ‘What Baboons Babies and Tetris Players Tell Us about Interaction.’ *Connection Science* 18(4), 363–378.

Dautenhahn, K. (2004) ‘Robots We Like to Live with? A developmental perspective on a personalized, life-long robot companion.’ in *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN 2004)*. held 20-22 September 2004 at Kurashiki, Okayama, Japan. New York, NY USA: IEEE, 17-22.

Demiris, Y. and Meltzoff, A. (2008) ‘The Robot in the Crib: A developmental analysis of imitation skills in infants and robots.’ *Infant and Child Development* 17(1), 43–53.

Dennett, D. (1989) *The Intentional Stance*. Cambridge, MA USA: MIT Press.

Dix, A., Finlay, J., Abowd, G.D. and Beale, R. (2004) *Human-Computer Interaction*. Harlow, Essex UK: Pearson Education Ltd.

Dominey, P.F. and Boucher, J-D. (2005) ‘Developmental Stages of Perception and Language Acquisition in a Perceptually Grounded Robot.’ *Cognitive Systems Research* 6(3), 243–259.

Esposito, A. (2009) ‘Affect in Multimodal Information.’ in *Affective Information Processing*. ed. by Tao, J. and Tan, T. London, UK: Springer London, 203–226.

Esposito, A. and Esposito, A. (2011) ‘On Speech and Gestures Synchrony.’ in Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud C. and Nijholt A. (eds.) *Analysis of Verbal and Nonverbal Communication and Enactment: COST2102 International Conference,* LNCS 6800. held 7-10 September 2010 at Budapest, Hungary. Berlin, Germany: Springer, 252–272.

Feldman, J.A. (2008) *From Molecules to Metaphor: A neural theory of language*. Cambridge, MA USA: Bradford Books.

Fitch, W.T. (2000) ‘The Evolution of Speech: A comparative review.’ *Trends in Cognitive Science* 4(7), 258–267.

Gales, M. and Young, S. (2007) ‘The Application of Hidden Markov Models in Speech Recognition.’ *Foundations and Trends in Signal Processing* 1(3), 195–304.

Gallese, V., Keysers, C. and Rizzolatti, G. (2004) ‘A Unifying View of the Basis of Social Cognition.’ *Trends in Cognitive Science* 8(9), 396–403.

Gasic, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P. and Young, S. (2013) ‘POMDP-Based Dialogue Manager Adaptation to Extended Domains.’ In Proceedings of the *Special Interest Group on Discourse and Dialogue (SIGDIAL)*. held 22-24 September 2013 in Metz, France. Stroudsburg, PA USA: ACL, 214-222.

Gates, W., Myhrvold, N. and Rinearson, P. (1995) *The Road Ahead*. New York, NY USA: Viking Penguin.

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M.P. and Mundell, A. (2005) ‘Designing Robots for Long-Term Social Interaction.’ in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*. held 2-5 August 2005 at Edmonton, Alberta, Canada. New York, NY USA: IEEE, 1338-1343.

Gorbunov, R., Barakova, E. and Rauterberg, M. (2013) ‘Design of Social Agents.’ *Neurocomputing* 114, 92–97.

Hastie, H., Lemon, O. and Dethlefs, N. (2012) ‘Incremental Spoken Dialogue Systems: Tools and data.’ in *Proceedings of NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community*. held 3-8 June 2012 at Montreal, Ontario, Canada. Stroudsburg, PA USA: ACL, 15-16.

Haugeland, J. (1985) *Artificial Intelligence: The very idea*. Cambridge, MA USA: MIT Press.

Hockett, C.F. (1960) ‘The Origin of Speech.’ *Scientific American* 203, 88–96.

Holden, C. (2004) ‘The Origin of Speech.’ *Science* 303, 1316–1319.

Jirak, D., Menz, M.M., Buccino, G., Borghi, A.M. and Binkofski, F. (2010) ‘Grasping Language - A short story on embodiment.’ *Consciousness and Cognition* 19(3), 711–720.

Kirby, R., Forlizzi, J. and Simmons, R. (2010) ‘Affective Social Robots.’ *Robotics and Autonomous Systems* 58(3), 322–332.

Kohler, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V. and Rizzolatti, G. (2002) ‘Hearing Sounds, Understanding Actions: Action representation in mirror neurons.’ *Science* 297, 846–848.

Komatsu, T. and Yamada, S. (2007) ‘How Appearance of Robotic Agents Affects How People Interpret the Agents’ Attitudes.’ in *International Conference on Advances in Computer Entertainment Technology*. held 13-15 June 2007 in Salzburg, Austria. New York, NY USA: ACM, 123-126.

Kurzweil, R. (1999) *The Age of Spiritual Machines*. New Haven, CT USA: Phoenix Press.

Kurzweil, R. (1990) *The Age of Intelligent Machines*. Cambridge, MA USA: MIT Press.

Lakatos, G., Gácsi, M., Topál, J. and Miklósi, Á. (2012) ‘Comprehension and Utilisation of Pointing Gestures and Gazing in Dog–Human Communication in Relatively Complex Situations.’ *Animal Cognition* 17(2), 201–213.

Lakoff, G. and Johnson, M. (1980), *Metaphors We Live By*. Chicago, IL USA: University of Chicago Press.

Lindblom, B. (1990) 'Explaining Phonetic Variation: A sketch of the H&H theory.' in *Speech Production and Speech Modelling*. ed. by Hardcastle, W.J. and Marchal, A. Dordrecht, The Netherlands: Kluwer Academic Publishers, 403–439.

Lombard, E. (1911) 'Le Sign de l'Elévation de la Voix.' *Annales Maladies Oreille, Larynx, Nez, Pharynx* 37, 101–119.

MacNeilage, P.F. (2008) *The Origin of Speech*. Cambridge, UK: Oxford University Press.

Marocco, D., Cangelosi, A., Fischer, K. and Belpaeme, T. (2010) 'Grounding Action Words in the Sensorimotor Interaction with the World: Experiments with a simulated iCub humanoid robot.' *Frontiers in Neurorobotics* 4(7).

Mehrabian, A. (1968) 'Communication without Words.' *Psychology Today* 2(9), 52–55.

Mitchell, W.J., Szerszen Sr., K.A., Lu, A.S., Schermerhorn, P.W., Scheutz, M. and MacDorman, K.F. (2011) 'A Mismatch in the Human Realism of Face and Voice Produces an Uncanny Valley.' *i-Perception* 2(1) 10–12.

Moore, R.K. (2013) 'Spoken Language Processing: Where do we go from here?' in *Your Virtual Butler*, LNAI 7407. ed. by Trappl R. Heidelberg, Germany: Springer, 111–125.

Moore, R.K. (2012) 'A Bayesian Explanation of the "Uncanny Valley" Effect and Related Psychological Phenomena.' *Nature Scientific Reports* 2 article 864. London, UK: Nature Publishing Group. available from < http://www.nature.com/srep/2012/121115/srep00864/full/srep00864.html>. [22 December 2013].

Moore, R.K. (2011) 'Interacting with Purpose (and Reeling!): What neuropsychology and the performing arts can tell us about "real" spoken language behaviour.' in Delgado, R.L.-C. and Kobayashi, T. (eds.) *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*. held 1-3 September 2011 at Berlin, Germany. Berlin, Germany: Springer, 5.

Moore, R.K. (2007a) 'Spoken language processing: piecing together the puzzle.' *Speech Communication* 49(5), 418–435.

Moore, R.K. (2007b) 'PRESENCE: A human-inspired architecture for speech-based human-machine interaction.' *IEEE Transactions on Computers* 56(9), 1176–1188.

Moore, R.K. (2007c) 'Towards speech-based human-robot interaction.' *Proceedings of the Symposium on Language and Robotics*. held 10-12 December, 2007 at Aveiro, Portugal. London, UK: AISB.

Moore, R.K. and Maier, V. (2012) 'Visual, vocal and behavioural affordances: some effects of consistency.' in *5th International Conference on Cognitive Systems - CogSys 2012*. held 23-24 February 2012 at Vienna, Austria. Vienna, Austria: Technical University of Vienna.

Moore, R.K. and Nicolao, M .(2011) 'Reactive Speech Synthesis: Actively managing phonetic contrast along an H&H continuum.' in *Proceedings of the 17th International Congress of Phonetics Sciences (ICPhS)*. held 17-21 August 2011 at Hong Kong, China. London, UK :IPA, 1422-1425.

Mori, M. (1970) ‘Bukimi no tani [The Uncanny Valley].’*Energy* 7(4), 33–35.

Nass, C. and Brave, S. (2005) *Wired for Speech: How voice activates and advances the human-computer relationship*. Cambridge, MA USA: MIT Press.

Newell, A., Barnett, J., Forgie, J.W., Green, C., Klatt, D., Licklider, J.C.R., Munson, J., Reddy, D.R. and Woods, W.A. (1973) *Speech Understanding Systems*. North Holland, The Netherlands: American Elsevier.

Nishio, S., Ishiguro, H. and Hagita, N. (2007) ‘Geminoid: Teleoperated android of an existing person.’ in *Humanoid Robots, New Developments*. ed. by de Pina Filho, A.C. Rijeka, Croatia: InTech, 343–352.

Otsuka, T., Nakadai, K., Takahashi, T., Komatani, K., Ogata, T. and Okuno, H.G. (2010) ‘Voice-Awareness Control for a Humanoid Robot Consistent with Its Body Posture and Movements.’ *Paladyn Journal of Behavioral Robotics* 1(1) 80–88.

Payr, S., Wallis, P., Cunningham, S. and Hawley, M. (2009) ‘Research on Social Engagement with a Rabbitic User Interface.’ [online] in *Adjunct Proceedings of the Ambient Intelligence: 3rd European Conference on Ambient Intelligence (AmI09),* LNCS 5859. held 18-21 November 2009 at Salzburg, Austria. Berlin-Heidelberg: Springer-Verlag. Available from L: < http://cogprints.org/6864/ >. [10 November 2013].

Pentland, A. (2008) *Honest Signals: how they shape our world*. London, UK: The MIT Press.

Picard, R. W. (1997) *Affective Computing*. Cambridge, MA USA: MIT Press.

Picheny, M. and Nahamoo, D. (2008) ‘Towards Superhuman Speech Recognition.’ in *Springer Handbook of Speech Processing SE –30*. ed. by Benesty, J., Sondhi, M.M. and Huang,Y. Berlin, Germany: Springer, 597–616.

Pickering, M.J. and Garrod, S. (2013) ‘An Integrated Theory of Language Production and Comprehension.’ *Behavioral and Brain Sciences* 36(4), 329–347.

Pieraccini, R. (2012) *The Voice in the Machine*. Cambridge, MA USA: MIT Press.

Premack, D. and Woodruff, G. (1978) ‘Does the Chimpanzee Have a Theory of Mind?’ *Behavioral and Brain Sciences* 1(4), 515–526.

Rabiner, L. (1997) open discussion session. *IEEE Workshop on Automatic Speech Recognition and Understanding*. held 17 December 1997 at Santa Barbara, CA USA.

Rizzolatti, G. and Craighero, L. (2004) ‘The Mirror-Neuron System.’ *Annual Review of Neuroscience* 27, 169–192.

Rizzolatti, G. and Arbib, M.A. (1998) ‘Language within Our Grasp.’ *Trends in Neuroscience* 21(5), 188–194.

Rogers, Y., Sharp, H. and Preece, J. (2011) *Interaction Design: Beyond human-computer interaction* (3rd ed.). Hoboken, NJ USA: John Wiley & Sons.

Scherer, K.R. (2003) ‘Vocal Communication of Emotion: A review of research paradigms.’ *Speech Communication* 40(1-2), 227–256.

Schmidhuber, J. (2006) ‘Developmental Robotics, Optimal Artificial Curiosity Creativity, Music, and the Fine Arts.’ *Connection Science* 18(2), 173–187.

Schwartz, J.L., Basirat, A., Ménard, L. and Sato, M. (2012) ‘The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception.’ *Journal of Neurolinguistics* 25(5), 336–354.

Serkhane, J.E., Schwartz, J.L. and Bessière, P. (2005) ‘Building a Talking Baby Robot: A contribution to the study of speech acquisition and evolution.’ *Interaction Studies* 6(2), 253–286.

Simon-Thomas, E.R., Keltner, D.J., Sauter, D., Sinicropi-Yao, L. and Abramson, A. (2009) ‘The Voice Conveys Specific Emotions: Evidence from vocal burst displays.’ *Emotion* 9(6), 838–846.

Skantze, G. and Schlangen, D. (2009) ‘Incremental Dialogue Processing in a Micro-Domain.’ in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09).* held 30 March – 3 April 2009 at Athens, Greece. Stroudsburg, PA USA: ACM, 745-753.

ten Bosch, L., Van hamme, H., Boves, L. and Moore, R.K. (2009) ‘A Computational Model of Language Acquisition: The emergence of words.’ *Fundamenta Informaticae* 90(3), 229–249.

Tomasello, M. (2008) *Origins of Human Communication*. Cambridge, MA USA: MIT Press.

Vernon, D. (2010) ‘Enaction As a Conceptual Framework for Developmental Cognitive Robotics.’ *Journal of Behavioral Robotics* 1(1), 89–98.

Vernon, D., Metta, G. and Sandini, G. (2010), ‘Embodiment in Cognitive Systems: On the mutual dependence of cognition and robotics.’ in *Advances in Cognitive Systems*. ed. by Gray, J. and Nefti-Meziani, S. Stevange, UK: Institution of Engineering and Technology (IET), 1-12.

Vinciarelli, A., Pantic, M. and Bourlard, H. (2009) ‘Social Signal Processing: Survey of an emerging domain.’ *Image and Vision Computing* 27(12), 1743–1759.

Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D’Errico, F. and Schröder, M. (2012) ‘Bridging the Gap between Social Animal and Unsocial Machine: A survey of social signal processing.’ *IEEE Transactions on Affective Computing* 3(1) 69–87.

Vogt, T., Andre, E. and Bee, N. (2008) ‘EmoVoice - A framework for online recognition of emotions from voice.’ in Delgado, R. L-C. and Kobayashi, T. (eds.) *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop,* LNCS 5078. Berlin-Heidelberg, Germany: Springer, 188–199.

Wagner, P., Malisz, Z. and Kopp, S. (2014) ‘Gesture and Speech in Interaction: An overview.’ *Speech Communication* 57, 209–232.

Walters, M., Syrdal, D., Koay, K., Dautenhahn, K. and te Boekhorst, R. (2008) ‘Human Approach Distances to a Mechanical-Looking Robot with Different Robot Voice Styles.’ *IEEE International Symposium on Robot and Human Interactive Communication*. held 1-3 August 2008 at Munich Germany, New York, NY USA: IEEE, 707-712.

Wilks, Y. (2010) *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*. Amsterdam, The Netherlands: John Benjamins Publishing Company.

Williams, J.D. and Young, S.J. (2007) ‘Partially Observable Markov Decision Processes for Spoken Dialog Systems.’ *Computer Speech and Language* 21(2), 231–422.

Winfield, A. (2012) *Robotics: A very short introduction*. Cambridge, UK: Oxford University Press.

Young, S.J. (2010) ‘Cognitive User Interfaces.’ *IEEE Signal Processing Magazine* 27(3), 128–140.

Zen, H., Tokuda, K. and Black, A.W. (2009) ‘Statistical Parametric Speech Synthesis.’ *Speech Communication* 51(11), 1039–1064.