Is Spoken Language All-or-Nothing? Implications for future speech-based human-machine interaction

Roger K. Moore

International Workshop on Spoken Dialogue Systems (IWSDS)
13-16 January 2016, Riekonlinna, Finland

Abstract Recent years have seen a significant market penetration of voice-based personal assistants (such as Apple's *Siri*) deployed on mobile devices. However, despite this success, user take-up is frustratingly low. This paper argues that there is a *habitability gap* caused by the inevitable mismatch between the capabilities and expectations of human users and the features provided by contemporary technical solutions. Some suggestions are made as to how such problems might be mitigated, but a more worrisome question surfaces: "*is spoken language all-or-nothing*"? Taking into consideration contemporary views on the special nature of (spoken) language, it is observed that there are situations where successful spoken language interaction can take place between mismatched interlocutors (such as between native and non-native speakers, or between an adult and a child, or even between a human being and a dog), and it is concluded that these scenarios might provide critical inspiration for the design of future speech-based human-machine interaction.

1 Introduction

The release, in 2011, of *Siri*, Apple's voice-based personal assistant for the iPhone signalled a step change in the public perception of spoken language technology. For the first time, a significant number of general users were exposed to the possibility of using their voice to enter information, navigate applications or pose questions - all by speaking to their mobile device. Of course, voice dictation software had been publicly available since the release of *Dragon Naturally Speaking* in 1997,

Roger K. Moore
Speech and Hearing Research Group, Dept. Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK, e-mail: r.k.moore@sheffield.ac.uk

but such technology only found success in niche market areas for document creation (by users who could not or would not type). In contrast, *Siri* appeared to offer a more general-purpose interface that thrust the potential benefits of automated speech-based interaction into the forefront of the public's imagination. By combining automatic speech recognition and speech synthesis with natural language processing and dialogue management, *Siri* promoted the possibility of a more *conversational* interaction between users and smart devices, and competitors such as Google's *Now* and Microsoft's *Cortana* soon followed[1].

Of course, it is well established that, while voice-based personal assistants such as Siri are now very familiar to the majority of mobile device users, their practical value is still in doubt. This is evidenced by the preponderance of videos on *YouTube* that depict humorous rather than practical uses; it seems that people give such systems a try, play around with them and then go back to their more familiar ways of doing things. Indeed, this has been confirmed by a recent survey of users from around the world which showed that only 13% used a facility such as *Siri* every day, whereas 46% had tried it once and then given up (citing inaccuracy and a lack of privacy as key reasons for abandoning it) [21].

This lack of serious take-up of voice-based personal assistants could be seen as the inevitable teething problems of a new(ish) technology, or it could be evidence of something more deep-seated. This *position* paper addresses these issues, and attempts to tease out some of the overlooked features of spoken language that might have a bearing on the success or failure of voice-based human-machine interaction. In particular, attention is drawn to the inevitable *mismatch* between the capabilities and expectations of human users and the features provided by contemporary technical solutions. Some suggestions are made as to how such problems might be mitigated, but a more worrisome question surfaces: "*is spoken language all-or-nothing*"?

2 The Nature of the Problem

There are many challenges facing the development of effective voice-based human-machine interaction. As the technology has matured, so the applications that are able to be supported have grown in depth and complexity (see Fig.1). From the earliest military *Command and Control Systems* to contemporary commercial *Interactive Voice Response (IVR) Systems* and the latest *Voice-Enabled Personal Assistants* (such as *Siri*), the variety of human accents, competing signals in the acoustic environment and the complexity of the application scenario have always presented significant barriers to practical usage [33]. Significant progress has been made in all of the core technologies, particularly following the emergence of the data-driven probabilistic modelling paradigm [10] (now supplemented by *deep learning* [15]) as a key driver in pushing regularly benchmarked performance in a positive direction.

[1] See [33] for a comprehensive review of the history of speech technology R&D up to, and including, the release of *Siri*.

Yet, as we have seen, usage remains a serious issue; not only does a speech interface compete with very effective non-speech GUIs [26], but people have a natural aversion to talking to machines in public spaces [21]. As Nass & Brave stated in their seminal book *Wired for Speech* [32]: "*voice interfaces have become notorious for fostering frustration and failure*" (p.6).

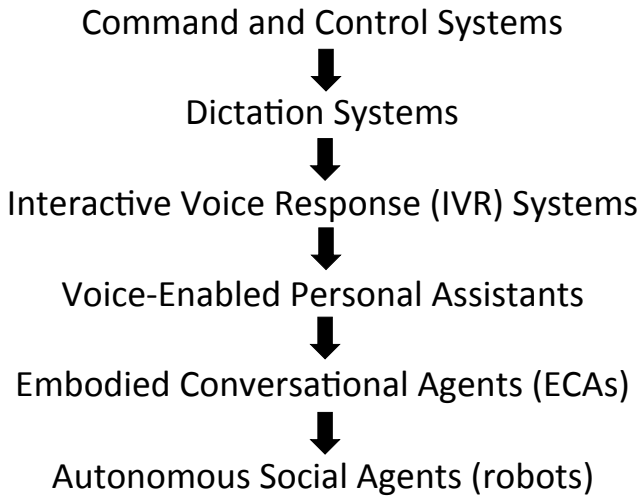

Fig. 1 The evolution of spoken language technology applications from early military *Command and Control Systems* to the much anticipated *Autonomous Social Agents* (robots).

These problems become magnified as the field moves forward to developing voice-based interaction with *Embodied Conversational Agents (ECAs)* and *Autonomous Social Agents* (robots). In these futuristic scenarios, it is assumed that spoken language will provide a "natural" conversational interface between human beings and so-called *intelligent* systems. However, there many challenges which need to be overcome in order to address such a requirement ...

> "*We need to move from developing robots that simply talk and listen to evolving intelligent communicative machines that are capable of truly understanding human behaviour, and this means that we need to look beyond speech, beyond words, beyond meaning, beyond communication, beyond dialogue and beyond one-off interactions.*" [30] (p.321)

Of these, a perennial problem seems to be how to evolve the complexity of voice-based interfaces from simple structured dialogues to more flexible conversational designs without confusing the user [2, 24, 22]. Indeed, it has been known for some time that there appears to be a non-linear relationship between *flexibility* and *usability* [35] - see Fig.2. As flexibility increases with advancing technology, so usability increases until a point where users no longer know what they can and cannot say, at which point usability tumbles and interaction falls apart.

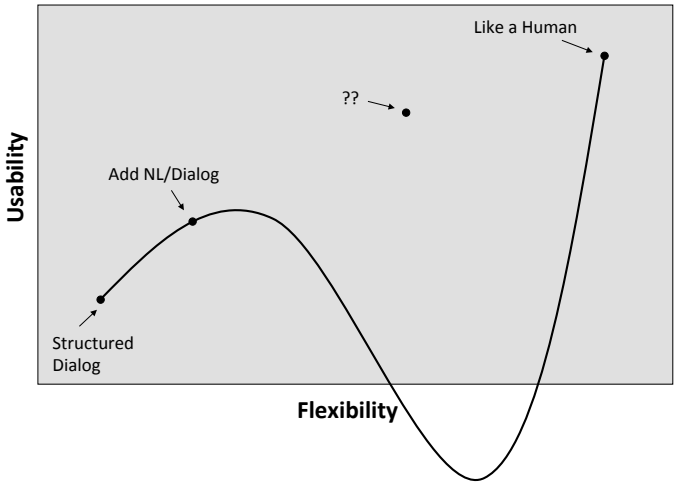

Fig. 2 Illustration of the consequences of increasing the flexibility of spoken language dialogue systems; increasing flexibility can lead to a *habitability gap* where usability drops catastrophically (graph reproduced, with permission, from Mike Phillips [35]). This means that it is surprisingly difficult to deliver a technology corresponding to the point marked as '??'. *Siri* sits at the point marked as 'Add NL/Dialog'.

2.1 The "Habitability Gap"

Progress is being made in this area: for example, by providing targeted help to users [39, 40, 18] and by replacing the traditional notion of turn-taking with a more fluid interaction based on *incremental processing* [37, 14]. Likewise, simple slot-filling approaches to language understanding and generation are being replaced by sophisticated statistical methods for estimating dialogue states and optimal next moves [41, 12]. Nevertheless, it is still the case that there is a *habitability gap* of the form illustrated in Fig.2.

In fact, the shape of the curve illustrated in Fig.2 is virtually identical to the famous *Uncanny Valley* effect in which a near human-looking artefact (such as a humanoid robot) can trigger feelings of eeriness and repulsion in an observer; as *human likeness* increases, so *affinity* increases until a point where the artefact starts to appear creepy [31]. A wide variety of explanations have been suggested for this non-linear relationship but, to date, there is only one quantitative model [28], and this is founded on the combined effect of categorical perception and mismatched perceptual cues. The implication of this model is that uncanniness - and hence, habitability - can be avoided if care is taken to align how an autonomous agent looks, sounds and behaves [29, 30]. In other words, if a speech-enabled agent is

to converse successfully with a human being, it should make clear its interactional *affordances* (in the sense defined originally by [13]).

This analysis leads to an important implication - since a spoken language system consists of a number of different components, each of which possesses a certain level of technical capability, then in order to be coherent (and hence *usable*), the design of the overall system needs to be aligned to the component with the *lowest* level of performance. For example, giving an automated personal assistant a natural human voice is a recipe for user confusion in the (normal) situation where the other speech technology components are limited in their abilities. In other words, in order to maximise the effectiveness of the interaction, a speech-enabled robot should have a robot voice[2]! This is an unpopular result[3], but there is evidence of its effectiveness [25], and it clearly has implications for contemporary voice-based personal assistants such as *Siri*, *Now* and *Cortana* which employ very humanlike voices[4].

2.2 Half a Language?

So far, so good - as component technologies improve, so the flexibility of the overall system would increase, and as long as the capabilities of the individual components are aligned, it should be possible to avoid falling into the habitability gap.

However, sending mixed messages about the capabilities of a spoken language system is only one part of the story; even if a speech-based autonomous social agent looks, sounds and behaves in a coherent way, will users actually be able to engage in conversational interaction if the overall capability is less than that normally enjoyed by a human being? What does it mean for a language-based system to be compromised in some way? How can users know what they may and may not say [17, 40], or even if this is the right question? Is there such a thing as *half* a language and, if so, is it habitable? Indeed, what is language anyway?

3 What is Language?

Unfortunately there is no space here to review the extensive and, at times, controversial history of the scientific study of language, or of the richness and variety of its spoken (and gestural) forms. Suffice to say that human beings have evolved a prolific system of (primary vocal) interactive behaviours that is vastly superior to that enjoyed by any other animal. As has been said many times ...

---

[2] As Balentine succinctly puts it: "*It's better to be a good machine than a bad person*" [1].

[3] It is often argued that such an approach is unimportant as users will habituate. However, habituation only occurs after sustained exposure, and a key issue here is how to increase the effectiveness of first encounters (since that has a direct impact on the likelihood of further usage).

[4] Interestingly, these ideas do appear to be having some impact on the design of contemporary autonomous social agents such as *Jibo* (which has a childlike and mildly robotic voice) [16].

"*Spoken language is the most sophisticated behaviour of the most complex organism in the known universe.*" [27].

The complexity and sophistication of (spoken) language tends to be masked by the apparent ease with which we use it. As a result, engineered solutions are often dominated by a somewhat naïve perspective involving the coding and decoding of messages passing from one brain (the sender) to another (the receiver). In reality, *languaging* is better viewed as an emergent property of the dynamic coupling between *cognitive unities* that serves to facilitate distributed sense-making through cooperative behaviours and, thereby, social structure [23, 5, 3, 4, 9]. Furthermore, the contemporary view is that language is based on the co-evolution of two key traits - *ostensive-inferential* communication and *recursive mind-reading* [38] - and that meaning is grounded in the physical world through *metaphor* [19, 6].

These modern perspectives on language not only place strong emphasis on *pragmatics* [20], but they are also founded on an implicit assumption that interlocutors are conspecifics[5] and hence share significant priors. Indeed, evidence suggests that some animals draw on representations of their own abilities (expressed as predictive models [8]) in order to interpret the behaviours of others [36, 42] and, for human beings, this is thought to be a key enabler for efficient recursive mind-reading [38] and hence for language [34, 11].

So now we arrive at an interesting position; if (spoken) language interaction between human beings is founded on shared experiences, representations and priors, to what extent is it possible to construct a technology that is intended to replace one of the participants? Could it be that there is a fundamental limit to the language-based interaction that can take place between *unequal* partners - between humans and machines? Indeed, returning to the question posed in Section 2.2 "*Is there such a thing as half a language?*", the answer appears to be "*no*"; spoken language appears to be all-or-nothing ...

"*The assumption of continuity between a fully coded communication system at one end, and language at the other, is simply not justified.*" [38] (p.46).

4 The Way Forward?

The story thus far provides a compelling explanation of the less-than-satisfactory experiences enjoyed by existing users of speech-enabled systems and identifies the source of the *habitability gap* outlined in Section 2.1. It would appear that, due to the gross mismatch between their respective priors, it might be impossible to create an automated system that would be capable of sustained and productive language-based interaction with a human being (except in narrow specialised domains involving experienced users). The vision of constructing a general-purpose voice-enabled autonomous social agent may be fundamentally flawed - the equivalent of trying to build a vehicle that travels faster than light!

[5] Members of the same species.

However, before we give up all hope, there are situations where voice-based interaction between mismatched partners is successful - but these are very different from the scenarios that are usually considered when designing current speech-based systems. For example, human beings regularly engage in vocal interaction with members of a different cultural and/or linguistic and/or generational background[6]. In such cases, all participants dynamically adjust many aspects of their behaviour - the clarity of their pronunciation, their choice of words and syntax, their style of delivery, etc. - all of which may be controlled by the perceived effectiveness of the interaction (that is, using *feedback* in a coupled system). Indeed, a particularly good example of such optimisation between mismatched interlocutors is the different way in which caregivers talk to children [7]. Maybe these same principles should be applied to speech-based human-machine interaction[7]?

Of course, these scenarios all involve spoken interaction between one human being and another, hence there is in reality a huge overlap of priors in terms of bodily morphology, environmental context and cognitive structure, as well as learnt social and cultural norms. Arguably the largest mismatch arises between an adult and a child, yet this is still interaction between conspecifics. A more extreme mismatch exists between non-conspecifics; for example, between humans and animals. However, it is interesting to note that our nearest relatives - the apes - do not have language, and this seems to be because they do not have ostensive communication (apes do not seem to understand pointing gestures) a key precursor to language.

Interestingly, one animal - the domestic dog - seems to excel in ostensive communication and, as a consequence, dogs are able to engage in very productive spoken language interaction with human partners (albeit one-sided and somewhat limited in scope) [38]. Spoken human-dog interaction is thus a potentially important example of a heavily mismatched yet highly effective cooperative configuration that might usefully inform spoken human-robot interaction in hitherto unanticipated ways.

5 Final Remarks

This paper has argued that there is a fundamental *habitability* problem facing contemporary spoken language systems, particularly as they penetrate the mass market and attempt to provide a general-purpose voice-based interface between human users and (so-called) intelligent systems. It has been suggested that the source of the difficulties in configuring genuinely usable systems is twofold: first, the need to align the visual, vocal and behavioural aspects of the system, and second, the need to overcome the huge mismatch between the capabilities and expectations of a hu-

[6] Interestingly, Nass & Brave noted that people speak to poor automatic speech recognition systems as if they were non-native listeners [32].

[7] Indeed, perhaps we should be explicitly studying the particular adaptations that human beings make when attempting to converse with autonomous social agents - a new variety of spoken language that could be appropriately termed "*robotese*"!

man being and those of even the most advanced autonomous social agent. This led to the preliminary conclusion that spoken language may indeed be all-or-nothing.

Of course, some might claim that the habitability problem only manifests itself in applications where task-completion is a critical measure of success. The suggestion would be that the situation might be different for applications in domains such as social robots, education or games in which the emphasis would be more on the spoken interaction itself. However, the argument presented in this paper is not concerned with the nature of the interaction, rather it questions whether such speech-based interaction can be sustained without access to the notion of *full* language.

Finally, and on a positive note, it was observed that there are situations where successful spoken language interaction can take place between mismatched interlocutors (such as between native and non-native speakers, or between an adult and a child, or even between a human being and a dog). It is thus concluded that these scenarios might provide critical inspiration for the design of future speech-based human-machine interaction.

Acknowledgements This work was supported by the European Commission [grant numbers EU-FP6-507422, EU-FP6-034434, EU-FP7-231868 and EU-FP7-611971], and the UK Engineering and Physical Sciences Research Council [grant number EP/I013512/1].

References

1. 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: ICMI Press.
2. Bernsen, N. O., Dybkjaer, H., Dybkjaer, L. (1998). *Designing Interactive Speech Systems: From First Ideas to User Testing*. Springer.
3. Bickhard, M. H. (2007). Language as an interaction system. *New Ideas in Psychology*, 25(2), 171-187.
4. Cowley, S. J. (Ed.). (2011). Distributed Language. John Benjamins Publishing Company.
5. Cummins, F. (2014). Voice, (inter-)subjectivity, and real time recurrent interaction. Frontiers in Psychology, 5, 760.
6. Feldman, J. A. (2008). *From Molecules to Metaphor: A Neural Theory of Language*. Bradford Books.
7. Fernald, A. (1985). Four-month-old infants prefer to listen to Motherese. *Infant Behavior and Development*, 8, 181-195.
8. Friston, K., Kiebel, S. (2009). Predictive coding under the free-energy principle. *Phil. Trans. R. Soc. B*, 364(1521), 1211-1221.
9. Fusaroli, R., Rączaszek-Leonardi, J., Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, 32, 147-157.
10. Gales, M., Young, S. J. (2007). The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3), 195-304.
11. Garrod, S., Gambi, C., Pickering, M. J. (2013). Prediction at all levels: forward model predictions can enhance comprehension. Language, *Cognition and Neuroscience*, 29(1), 46-48.
12. Gasic, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P., Young, S. J. (2013). POMDP-based dialogue manager adaptation to extended domains. In *SIGDIAL* (pp. 214-222). Metz, France.

13. Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (pp. 67-82). Hillsdale, NJ: Lawrence Erlbaum.
14. Hastie, H., Lemon, O., 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* (pp. 15-16). Montreal, Canada.
15. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *Signal Processing Magazine*, IEEE.
16. *Jibo: The World's First Social Robot for the Home*, https://www.jibo.com
17. Jokinen, K., Hurtig, T. (2006). User expectations and real experience on a multimodal interactive system. In *INTERSPEECH-ICSLP Ninth International Conference on Spoken Language Processing*. Pittsburgh, PA, USA.
18. Komatani, K., Fukubayashi, Y., Ogata, T., Okuno, H. G. (2007). Introducing utterance verification in spoken dialogue system to improve dynamic Help generation for novice users. In *8th SIGdial Workshop on Discourse and Dialogue* (pp. 202-205).
19. Lakoff, G., Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.
20. Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
21. Liao, S.-H. (2015). *Awareness and Usage of Speech Technology*. Masters thesis, Dept. Computer Science, University of Sheffield.
22. Lopez Cozar Delgado, R. (2005). *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment*. Wiley.
23. Maturana, H. R., Varela, F. J. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston, MA: New Science Library/Shambhala Publications.
24. McTear, M. F. (2004). *Spoken Dialogue Technology: Towards the Conversational User Interface*. Springer.
25. Moore, R. K., Morris, A. (1992). Experiences collecting genuine spoken enquiries using WOZ techniques. *5th DARPA Workshop on Speech and Natural Language*. New York.
26. Moore, R. K. (2004). Modelling data entry rates for ASR and alternative input methods. In *INTERSPEECH-ICSLP*. Jeju, Korea.
27. Moore, R. K. (2007). Spoken language processing: piecing together the puzzle. *Speech Communication*, 49(5), 418-435.
28. Moore, R. K. (2012). A Bayesian explanation of the "Uncanny Valley" effect and related psychological phenomena. *Nature Scientific Reports*, 2(864).
29. Moore, R. K., Maier, V. (2012). Visual, vocal and behavioural affordances: some effects of consistency. *5th International Conference on Cognitive Systems (CogSys 2012)*. Vienna.
30. Moore, R. K. (2015). From talking and listening robots to intelligent communicative machines. In J. Markowitz (Ed.), *Robots That Talk and Listen* (pp. 317-335). Boston, MA: De Gruyter.
31. Mori, M. (1970). Bukimi no tani (the uncanny valley). *Energy*, 7, 33-35.
32. Nass, C., Brave, S. (2005). *Wired for Speech: How Voice Activates and Advances the Human-computer Relationship*. Cambridge, MA: MIT Press.
33. Pieraccini, R. (2012). *The Voice in the Machine*. MIT Press, Cambridge, MA.
34. Pickering, M. J., Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, 11(3), 105-110.
35. Philips, M. (2006). Applications of spoken language technology and systems. In M. Gilbert & H. Ney (Eds.), *IEEE/ACL Workshop on Spoken Language Technology (SLT)*.
36. Rizzolatti, G., Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169-192.
37. Schlangen, D., Skantze, G. (2009). A general, abstract model of incremental dialogue processing. *12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*. Athens, Greece.
38. Scott-Phillips, T. (2015). *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Palgrave MacMillan.

39. Tomko, S., Harris, T. K., Toth, A., Sanders, J., Rudnicky, A., Rosenfeld, R. (2005). Towards efficient human machine speech communication. *ACM Transactions on Speech and Language Processing*, 2(1), 1-27.
40. Tomko, S. L. (2006). *Improving User Interaction with Spoken Dialog Systems via Shaping*. PhD Thesis, Carnegie Mellon University.
41. Williams, J. D., Young, S. J. (2007). Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2), 231-422.
42. Wilson, M., Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131(3), 460-473.