A Needs-Driven Cognitive Architecture for Future 'Intelligent' Communicative Agents

Roger K. Moore
Dept. Computer Science, University of Sheffield, UK
Email: r.k.moore@sheffield.ac.uk

***Abstract*—Recent years have seen considerable progress in the deployment of 'intelligent' communicative agents such as Apple's *Siri*, Google *Now*, Microsoft's *Cortana* and Amazon's *Alexa*. Such speech-enabled assistants are distinguished from the previous generation of voice-based systems in that they claim to offer access to services and information via conversational interaction. In reality, interaction has limited depth and, after initial enthusiasm, users revert to more traditional interface technologies. This paper argues that the standard architecture for a contemporary communicative agent fails to capture the fundamental properties of human spoken language. So an alternative needs-driven cognitive architecture is proposed which models speech-based interaction as an emergent property of coupled hierarchical feedback control processes. The implications for future spoken language systems are discussed.**

I. Introduction

The performance of spoken language systems has improved significantly in recent years, with corporate giants such as MicroSoft and IBM issuing claim and counter-claim as to who has the lowest word error rates. Such progress has contributed to the deployment of ever more sophisticated voice-based applications, from the earliest military 'Command and Control Systems' to the latest consumer 'Voice-Enabled Personal Assistants' (such as *Siri*) [1]. Research is now focussed on voice-based interaction with 'Embodied Conversational Agents (ECAs)' and 'Autonomous Social Agents' based on the assumption that spoken language will provide a 'natural' conversational interface between human beings and future (so-called) *intelligent* systems – see Fig. 1.

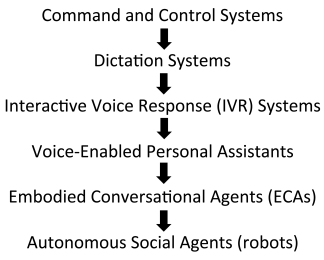

Fig. 1. The evolution of spoken language technology applications.

In reality, users' experiences with contemporary spoken language systems leaves a lot to be desired. After initial enthusiasm, users lose interest in talking to *Siri* or *Alexa*, and they revert to more traditional interface technologies [2]. One possible explanation for this state of affairs is that, while component technologies such as automatic speech recognition and text-to-speech synthesis are subject to continuous ongoing improvement, the overall architecture of a spoken language system has been standardised for some time [3] – see Fig. 2. Standardisation is helpful because it promotes interoperability and expands markets. However, it can also stifle innovation by prescribing sub-optimal solutions. So, what (if anything) might be wrong with the architecture illustrated in Fig. 2?

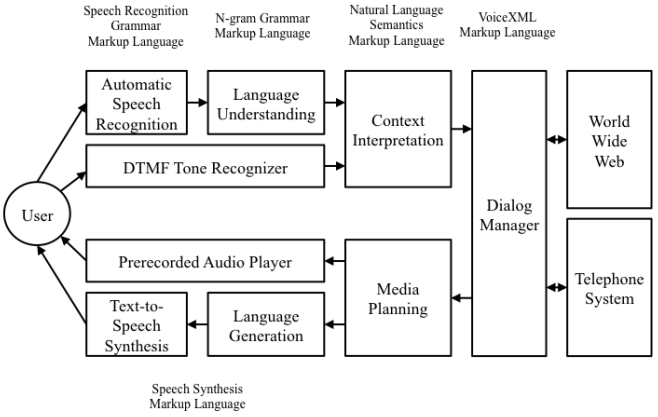

Fig. 2. Illustration of the W3C *Speech Interface Framework* [3].

In the context of spoken language, the main issue with the architecture illustrated in Fig. 2 is that it reflects a traditional stimulus–response ('behaviourist') view of interaction; the user utters a request, the system replies. This is the 'tennis match' analogy for language; a stance that is now regarded as restrictive and old-fashioned. Contemporary perspectives regard spoken language interaction as being more like a three-legged race than a tennis match [4]: continuous coordinated behaviour between coupled dynamical systems.

II. Towards a 'Cognitive' Architecture

What seems to be required is an architecture that replaces the traditional 'open-loop' stimulus-response arrangement with a 'closed-loop' dynamical framework; a framework in which needs/intentions lead to actions, actions lead to consequences, and perceived consequences are compared to intentions/needs (in a continuous cycle of synchronous

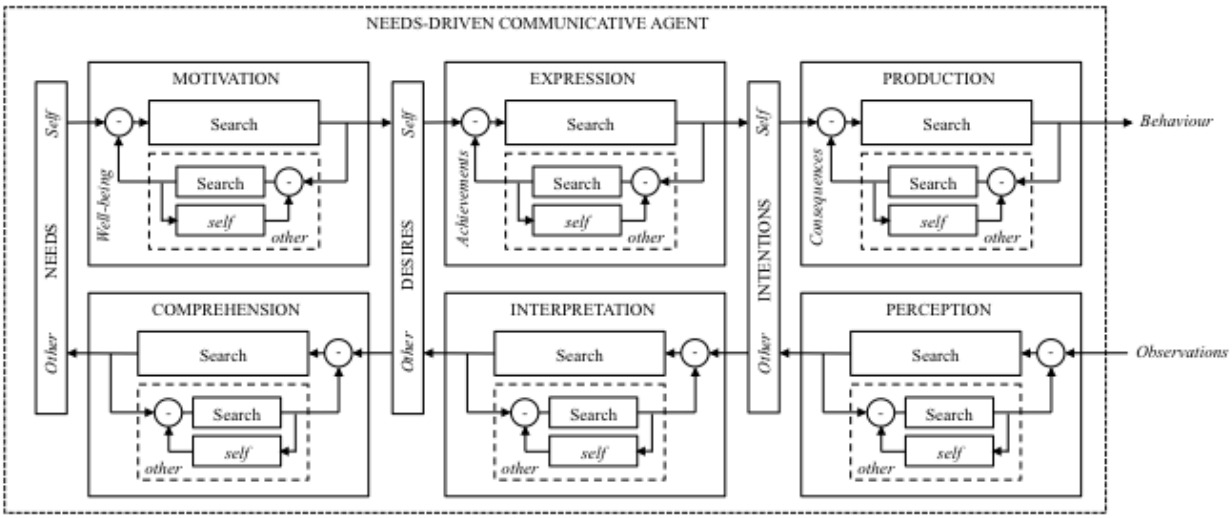

Fig. 3. Illustration of the proposed architecture for a needs-driven communicative agent [7].

behaviours). Such an architecture has been proposed by the author [5], [6], [7] – see Fig. 3.

One of the key concepts embedded in the architecture illustrated in Fig. 3 is the agent's ability to 'infer' (using search) the consequences of their actions when they cannot be observed directly. Another is the use of a *forward model* of 'self' to model 'other'. Both of these features align well with the contemporary view of language as "*ostensive inferential recursive mind-reading*" [8]. Also, the architecture makes an analogy between the depth of each search process and 'motivation/effort'. This is because it has been known for some time that speakers continuously trade effort against intelligibility [9], [10], and this maps very nicely into a hierarchical control-feedback process [11] which is capable of maintaining sufficient contrast at the highest *pragmatic* level of communication by means of suitable regulatory compensations at the lower semantic, syntactic, lexical, phonemic, phonetic and acoustic levels.

As a practical example, these ideas have been used to construct a new type of speech synthesiser (known as 'C2H') that adjusts its output as a function of its inferred communicative success [13], [14] – it listens to itself!

III. Final Remarks

Whilst the proposed cognitive architecture successfully captures some of the key elements of language-based interaction, it is important to note that such interaction between human beings is founded on substantial shared priors. This means that there may be a fundamental limit to the language-based interaction that can take place between *mismatched* partners such as a human being and an autonomous social agent [15].

Acknowledgment

This work was partially supported by the European Commission [EU-FP6-507422, EU-FP6-034434, EU-FP7-231868 and EU-FP7-611971], and the UK Engineering and Physical Sciences Research Council (EPSRC) [EP/I013512/1].

References

[1] R. Pieraccini. *The Voice in the Machine*. MIT Press, Cambridge, 2012.
[2] R. K. Moore, H. Li, & S.-H. Liao. Progress and prospects for spoken language technology: what ordinary people think. In *INTERSPEECH* (pp. 3007–3011). San Francisco, CA, 2016.
[3] Introduction and Overview of W3C Speech Interface Framework, http://www.w3.org/TR/voice-intro/
[4] F. Cummins. Periodic and aperiodic synchronization in skilled action. *Frontiers in Human Neuroscience*, 5(170), 1–9, 2011.
[5] R. K. Moore. PRESENCE: A human-inspired architecture for speech-based human-machine interaction. *IEEE Trans. Computers*, 56(9), 1176–1188, 2007.
[6] R. K. Moore. Spoken language processing: time to look outside? In L. Besacier, A.-H. Dediu, & C. Martn-Vide (Eds.), *2nd International Conference on Statistical Language and Speech Processing (SLSP 2014), Lecture Notes in Computer Science* (Vol. 8791). Springer, 2014.
[7] R. K. Moore. PCT and Beyond: Towards a Computational Framework for "Intelligent" Systems. In A. McElhone & W. Mansell (Eds.), *Living Control Systems IV: Perceptual Control Theory and the Future of the Life and Social Sciences*. Benchmark Publications Inc. In Press (available at https://arxiv.org/abs/1611.05379).
[8] T. Scott-Phillips. *Speaking Our Minds: Why human communication is different, and how language evolved to make it special.* London, New York: Palgrave MacMillan, 2015.
[9] E. Lombard. Le sign de l?lvation de la voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37, 101–119, 1911.
[10] B. Lindblom. Explaining phonetic variation: a sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling* (pp. 403–439). Kluwer Academic Publishers, 1990.
[11] W. T. Powers. *Behavior: The Control of Perception*. NY: Aldine: Hawthorne, 1973.
[12] S. Hawkins. Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, 31, 373–405, 2003.
[13] R. K. Moore & M. Nicolao. Reactive speech synthesis: actively managing phonetic contrast along an H&H continuum, *17th International Congress of Phonetics Sciences (ICPhS)*. Hong Kong, 2011.
[14] M. Nicolao, J. Latorre & R. K. Moore. C2H: A computational model of H&H-based phonetic contrast in synthetic speech. *INTERSPEECH*. Portland, USA, 2012.
[15] R. K. Moore. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In K. Jokinen & G. Wilcock (Eds.), *Dialogues with Social Robots - Enablements, Analyses, and Evaluation.* Springer Lecture Notes in Electrical Engineering, 2016.