Towards Speech-Based Human-Robot Interaction

Roger K. Moore

Dept. Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK
r.k.moore@dcs.shef.ac.uk

Abstract

Notwithstanding the success of contemporary spoken language technology in a range of practical applications, it is widely acknowledged that serious shortfalls in performance limit its wider deployment. Unconstrained speech-based interaction with embodied agents - such as robots - remains outside the scope of current technology and thus presents key challenges to the research community. However, it is argued that the solutions lie, not only outside the field of speech technology, but also outside current theories of human spoken language processing. Instead, it is proposed that research into spoken language by mind *or* machine now needs to draw inspiration from areas as widely dispersed as cognitive neuroscience and control engineering. Following such an approach, this paper describes a theoretical framework known as 'PREdictive SENsorimotor Control and Emulation' (PRESENCE), and experiments using a PRESENCE-inspired architecture to enable a robot to clap in synchrony with a user's voice illustrate the power of the paradigm. It is concluded that future research in spoken language processing is likely to benefit greatly from PRESENCE and from greater emphasis on the challenges raised in situated and embodied environments, the evolution and acquisition of spoken language, and appropriate and intuitive speech-based human-robot interaction.

Introduction

Over the past fifty years, spoken language technology – automatic speech recognition, text-to-speech synthesis and spoken language dialogue systems – has made tremendous strides in terms of its technical abilities and practical applications. The majority of mobile telephones now carry 'voice dialling' as a standard feature, the new Microsoft Vista operating system incorporates the ability to dictate documents or control a PC by voice, and IVR (interactive voice response) systems are becoming commonplace for interacting with automated services over the telephone. Progress has been driven by the extensive use of machine learning techniques drawing on vast quantities of speech training material.

However, these successes belie the uncomfortable fact that the performance of such systems appears to be asymptoting well short of human spoken language capabilities, and such shortfalls reveal themselves in realistic everyday environments which may contain competing sound sources, multiple users or which inadvertently encourages users to step outside the narrow confines of the application domain. Unfortunately each of these aspects typifies the range of applications that involve speech-based interaction with embodied agents - such as robots - and hence the feasibility of integrating contemporary spoken language technology into robotic systems is currently severely compromised.

Nevertheless, the challenges posed by attempting to speech-enable robotic systems are exactly those that can drive spoken language technology research in fruitful new directions. The author has argued elsewhere (Moore, 2007a) that the limitations of current spoken language technology are a direct consequence of the natural tendency of scientists to take a reductionist approach in which automatic speech recognition, synthesis and dialogue are treated as independent components and even developed by different research communities. Such enforced separation also undermines those few attempts that have been made to 'bridge the gap' between automatic and human speech processing (Scharenborg et al, 2003).

The Way Forward

What appears to be needed to move to the next generation of spoken language technology is to re-evaluate the current research paradigms not, as one might suppose, with respect to current theories of human spoken language (which are similarly fragmented), but in the light of a number of advanced ideas drawn from disciplines *outside* the field of spoken language processing. In particular, considerable progress is currently being made (in areas such as cognitive neuroscience) in understanding and modelling the general behaviour of living systems, and much of this research is directly relevant to spoken language interaction. Old ideas such as 'perceptual control theory' (Powers, 1973) and new discoveries such as 'mirror neurons' (Rizzolatti and Craighero, 2004) serve to indicate a hitherto unsuspected and intimate link between perceptual and productive behaviours and inspire new models of action understanding based on significant sensorimotor overlap. Coupled with contemporary theories of cortical functionality such as 'hierarchical temporal memory' (Hawkins, 2004) and 'emulators' (Grush, 2004), these putative processes offer a tantalising glimpse into possible computational models of cognition, interaction and speech.

Predictive Sensorimotor Control and Emulation

In (Moore, 2007a and 2007b), the author has drawn a number of such ideas together into a single coherent theoretical framework termed PRESENCE – 'PREdictive SENsorimotor Control and Emulation' - a core feature of which is the necessity to move away from a classic

Brunswikian stimulus-response model of behaviour to one in which participants (humans or machines) are viewed as multiple loosely-coupled control-feedback loops. It is argued that such an approach provides a more sophisticated model of interactive behaviour such as spoken language and provides a putative architecture for future speech-based human-machine interaction in situated embodied environments.

PRESENCE is based on the premise that there are three fundamental factors that ultimately determine an organism's fitness to survive in an evolutionary framework: its ability to manage **energy** (facilitating efficient behaviour in the context of scarce resources), **time** (facilitating efficient planning in the context of potentially harmful situations) and **entropy** (facilitating efficient communications in the context of information sparsity). These constraints, coupled with an integrated and recursive processing architecture, pave the way to a new approach to spoken language technology in which high-level interactive behaviours such as prosody and emotion emerge as essential aspects of a communicative system rather than as processing afterthoughts.

Experimental Work

A preliminary experimental validation of the principles espoused in PRESENCE has been conducted using the ALPHA REX humanoid robot constructed using the LEGO® MINDSTORMS® NXT platform. By coordination and synchronization in a PRESENCE-inspired framework, the robot was able to learn to produce motor behaviour in time to rhythmic spoken input (much like someone clapping along to music).

The robot was programmed using three sensorimotor control loops: one to monitor and control its own behaviour, one to monitor the behaviour of the human user and a third driven by a 'need' to optimise synchronisation between the other two. The resulting behaviour is illustrated in figure 1.

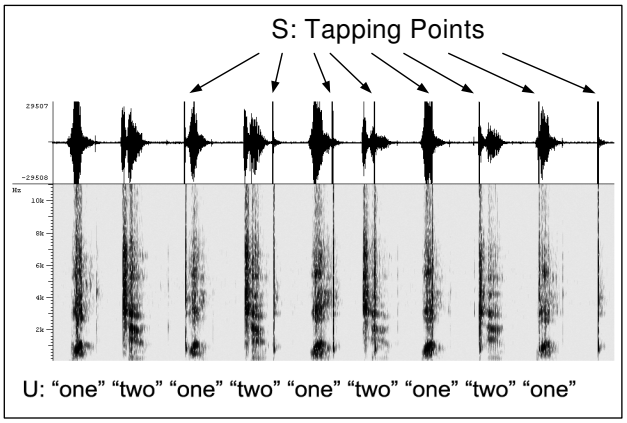

Fig.1: Robot tapping in synchrony with a user's speech (S-system, U-user).

The results of the experiment showed that the robot was not only able to synchronise its behaviour with that of the user, but it also successfully predicted successive rhythmic actions after the user ceased to speak.

Conclusion & Future Research

As a result of the development of PRESENCE and the preliminary experiments reported here, it is concluded that future research in spoken language processing is likely to benefit greatly from greater emphasis on the very practical issues raised in situated and embodied environments, and the computational mechanisms required to support appropriate and intuitive speech-based human-robot interaction. To that end, research at the University of Sheffield is currently being directed towards models of the evolution and acquisition of spoken language (Boves et al, 2007), and the development of an animatronic tongue - see figure 2.

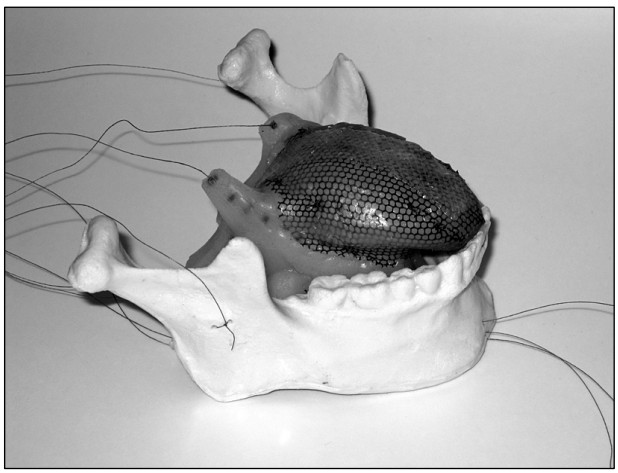

Fig. 2: Animatronic tongue being developed at the University of Sheffield.

References

Boves L., ten Bosch L. and Moore R.K. 2007. ACORNS: Towards computational modeling of communication and recognition skills, Proc. 6th IEEE Int. Conf. on Cognitive Informatics, Lake Tahoe, CA, USA.

Grush, R. 2004. The emulation theory of representation: motor control, imagery, and perception, Behavioral and Brain Sciences 27:377-442.

Hawkins, J. 2004. On Intelligence, Times Books.

Moore, R.K. 2007a. Spoken language processing: piecing together the puzzle, Speech Communication 49:418-435.

Moore, R.K. 2007b. PRESENCE: A human-inspired architecture for speech-based human-machine interaction, IEEE Trans. Computers, 56:1176-1188.

Powers, W.T. 1973. Behaviour: The Control of Perception, Hawthorne, NY: Aldine.

Rizzolatti, G. and Craighero, L. 2004. The mirror-neuron system, Annual Review of Neuroscience 27:169-192.

Scharenborg, O., ten Bosch, L., Boves, L. and Norris, D. 2003. Bridging automatic speech recognition and psycholinguistics: extending Shortlist to an end-to-end model of human speech recognition, J. Acoustical Soc. of America 114(6):3023-3035.