In 2nd International Conference on Statistical Language and Speech Processing (SLSP 2014). Grenoble.

Spoken Language Processing: Time to Look Outside?

Roger K. Moore

Speech and Hearing Research Group, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
http://www.dcs.shef.ac.uk/~roger
r.k.moore@sheffield.ac.uk

Abstract. Over the past thirty years, the field of spoken language processing has made impressive progress from simple laboratory demonstrations to mainstream consumer products. However, commercial applications such as *Siri* highlight the fact that there is still some way to go in creating *Autonomous Social Agents* that are truly capable of conversing effectively with their human counterparts in real-world situations. This paper suggests that it may be time for the spoken language processing community to take an interest in the potentially important developments that are occurring in related fields such as cognitive neuroscience, intelligent systems and developmental robotics. It then gives an insight into how such ideas might be integrated into a novel *Mutual Beliefs Desires Intentions Actions and Consequences* (MBDIAC) framework that places a focus on generative models of communicative behaviour which are recruited for interpreting the behaviour of others.

Keywords: spoken language processing, enactivism, language grounding, mirror neurons, perceptual control, cognitive architectures, autonomous social agents

1 Introduction

Since the 1980s, the introduction of stochastic modelling techniques - particularly hidden Markov models (HMMs) - into the field of spoken language processing has given rise to steady year-on-year improvements in capability [1], [2]. Coupled with a relentless increase in the processing power of the necessary computing infrastructure, together with the introduction of public benchmark testing, the field has developed from a specialist area of engineering research into the commercial deployment of mainstream consumer products. With the advent of smartphone applications such as Apple's Siri, Microsoft's Cortana and Google's Now, speech-based interaction with 'intelligent' devices has entered the popular imagination, and public awareness of the potential benefits of hands-free access to information is at an all-time high [3].

The gains in performance for component technologies such as automatic speech recognition and text-to-speech synthesis have accrued directly from the deployment of state-of-the-art machine learning techniques in which significantly large corpora of annotated speech (often thousands of hours) are used to estimate the parameters of rich context-sensitive Bayesian models. Indeed, the immense challenges posed by the need to create accurate and effective spoken language processing has meant that speech technology researchers have become acknowledged pioneers in the use of the most advanced machine learning techniques available. A recent example of this is the performance gains arising from the use of deep neural networks (DNNs) [4].

However, notwithstanding the immense progress that has been made over the past thirty or so years, it is generally acknowledged that there is still some way to go before spoken language technology systems are sufficiently reliable for the majority of envisaged applications. Whilst the performance of state-of-the-art systems is impressive, it is still well short of what is required to provide users with an effective and reliable alternative to traditional interface technologies such as keyboards and touch-sensitive screens [5]. Moreover, it is clear that spoken language capabilities of the average human speaker/listener are considerably more robust in adverse real-world situations such as noisy environments, dealing with speakers with foreign accents or conversing about entirely novel topics. This means that there is still a clear need for significant improvements in our ability to model and process speech, and hence it is necessary to ask where these gains might arise - more training data, better models, new algorithms, or from some other source [6]?

It is posited here that it is time for the spoken language processing community to look outside the relatively narrow confines of the discipline in order to understand the potentially important developments that are taking place in related areas. Fields such as cognitive neuroscience, intelligent systems and developmental robotics are progressing at an immense pace and, although some of the tools and techniques employed in spoken language processing could be of value to those fields, there is a growing understanding outside the speech area of how living systems are organised and how they interact with the world and with each other. Some of these new ideas could have a direct bearing on future spoken language systems, and could provide a launchpad for the kinds of developments that are essential if the potential of speech-based language interaction with machines is to be realised fully. This paper addresses these issues and introduces a number of key ideas from outside the field of spoken language processing which the author believes could be of some significance to future progress.

2 Looking for Inspiration Outside

It is often remarked that spoken language could be the most sophisticated behaviour of the most complex organism we know [7], [8], [9]. However, the apparent ease with which we as human beings interact using speech tends to mask the variety and richness of the mechanisms that underpin it. In fact the spoken language processing research community has become so focused on the rather obvious surface patterning - such as lexical structure (i.e. words) - that the foundational principles on which spoken *interaction* is organised has a tendency to be overlooked. In reality, long before spoken language dialogue evolved as a rich communicative behaviour, the distant ancestors of modern human beings were coordinating their activities using a variety of communicative modes and behaviours (such as the synchronisation of body postures, making explicit gestures, the laying down of markers in the environment and the use of appropriate sounds and noises). Interactivity is thus a fundamental aspect of the behaviour of living systems, and it would seem appropriate to found spoken language interaction on more primitive behaviours.

Interestingly, interactivity is not solely concerned with the behavioural relationship between one organism and another. In the general case, interactivity takes place between an organism and its physical *environment*, where that environment potentially incorporates other living systems. From an evolutionary perspective, interactive behaviour between an organism and its environment can be seen to emerge as a survival mechanism aimed at maintaining the persistence of an organism long enough for successful procreation, and these are issues that have engaged deep thinking theorists for any years. Of particular relevance here is the growth of an approach to understanding (and modelling) living systems known as *enactivism*.

2.1 Enactivism

Enactivism grew out of seminal work by Humberto Maturana and Francisco Varela [10] in which they tackled fundamental questions about the nature of living systems. In particular, they identified *autopoiesis* (a process whereby organisational structure is preserved over time) as a critical self-regulatory mechanism and *cognition* (the operation of a nervous system) as providing a more powerful and flexible autopoietic mechanism than purely chemical interactions. They defined a minimal living system such as a single cell as an *autopoietic unity* whereby the cell membrane maintains the boundary between the unity and everything else. Hence, a unity is said to be structurally *coupled* with its external environment - 1st-order coupling - and, for survival, appropriate interactive behaviours are required to take place (such as moving up a sugar gradient).

Likewise, unities may be coupled with other unities forming *symbiotic* or *metacellular* organisational structures - 2nd -order coupling - which can then be viewed organisationally as unities in their own right. The neuron is cited as a special type of cell emerging from particular symbiotic coupling, and the nervous system is thus seen as facilitating a special form of 2nd-order metacellular organisation termed a *cognitive unity.* Finally, Maturana and Varela propose that interaction between cognitive unities - 3rd-order coupling - is manifest in the form of organised *social systems* of group behaviour, and the emergence of cooperation, communication and language are posited as a consequence of 3rd-order coupling.

The enactive perspective thus establishes a simple and yet powerful framework for understanding the complexity of interaction between living systems, and it holds the promise for the investigation of computational approaches that seek to mimic these same behaviours. The emphasis on the coupling between a cognitive unity and its external environment (including other unities) is central to the approach, and this provides two clear messages - interactivity must be viewed as essentially *multimodal* in nature and that interaction is *grounded* in the context in which it takes place. Likewise, enactivism makes it clear that (spoken) language interaction is founded upon more general-purpose behaviours for continuous communicative coupling rather than simple turn-by-turn message passing [11], [12].

2.2 Multimodal Interaction and Communication

In principle, the modality in which interaction between a living system and its environment (including other living systems) takes place should be irrelevant. However, in practice, the characteristics and affordances [13] of the different modes greatly influence the modalities employed. For example, it may be easier to move a heavy object by pushing it bodily rather than by blowing air at it. Similarly, it may be safer to influence the behaviour of another living system by making a loud noise from a distance rather than by approaching it and touching it physically.

Nevertheless, notwithstanding the static advantages and disadvantages of any particular mode of interaction, in a dynamic and changing world it makes sense for an organism to be able to actively distribute information across alternative modes as a function of the situational context. Hence, even a sophisticated behaviour such as language should be viewed as being essentially a multimodal activity. Given this perspective, it would be natural to assume that there exists some significant relationship between physical gestures and vocal sounds. In such a framework, the power of multimodal behaviour such as speaking and pointing would be taken for granted, and the emergence of prosody as a fundamental carrier of unimodal vocal pointing behaviour would be more obvious.

For an up-to-date review of multimodal integration in general, see [14], and for speech and gesture in particular, see [15]. The argument here is that such behaviours are not simply 'nice to have' additional features (as they tend to be treated currently), but that they represent the basic substrate on which spoken language interaction is founded. Indeed a number of authors have argued that vocal language evolved from gestural communication (freeing up the hands for tool use or grooming) [16], [17], [18]. Hence, these insights suggest that information about multimodal characteristics and affordances should be intrinsic to the computational modelling paradigms employed in spoken language systems.

2.3 Language Grounding

The notion that an organism is not only coupled with its environment, but also with other organisms in the environment, introduces another important and fundamental aspect of interactive behaviour - passive information flow versus active signalling. In the first case, almost any behaviour could have indirect consequences in the sense that the environment could be disturbed by any physical activity, and such disturbance may provide a cue to other organisms as to what has taken place. As a result, organisms could exploit the availability of such information for their own benefit; for example, a predator could track a prey by following its trail of scent. In this situation, the emergent coupled behaviour is conditioned upon passive (unintentional) information transfer between individual organisms via the environment. In this case, the information laid down in the environment has *meaning* for the receiver, but not for the sender. However, living systems may also actively manage the information flow, and this would take the form of active (intentional) signalling - signals that have meaning for the sender (and hopefully, the receiver). In fact the value of *shared* meanings is immense, and

Meaning and semantics have been rather latecomers to the spoken language processing party. However, from the perspective being developed here, it is clear that the significance and implications of a behaviour are fundamental to the dynamics of the coupling that takes place between one individual and another. In other words, meaning is everything! The implication of this view is that the coupling is contingent on the *communicative context* which, in general terms, consists of the characteristics of the agents involved, the physical environment in which they are placed and the temporal context in which the actions occur. In modern terminology, meaningful communication is said to be *grounded* in the real world [19], and that generating and interpreting such behaviour is only possible with reference to the *embodied* nature and *situated* context in which the interactions take place. The grounding provided by a common physical environment gives rise to the possibility of *shared* meanings and representations [20], and crucial behaviours such *joint attention* and *joint action* emerge as a direct consequence of managing the interaction [21], [22], [23], [24].

Such a perspective has taken strong hold in the area of developmental robotics in which autonomous agents acquire communication and language skills (and in particular, meanings) not through instruction, but through interaction [25], [26], [27], [28], [29]. These approaches address the *symbol grounding problem* [30] by demonstrating that linguistic structure can be mapped to physical movement and sensory perception. As such, they represent the first steps towards a more general approach which hypothesises that even the most abstract linguistic expressions may be understood by the use of *metaphor* to link high-level representations to low-level perceptions and actions [31].

2.4 Mirror Neurons and Simulation

One of the drivers behind grounding language in behaviour is the discovery in the 1990s of a neural mechanism - so-called *mirror neurons* - that links action and perception [32], [33]. The original experiment involved the study of neural activity in the motor cortex of a monkey grasping a small item (such as a raisin). The unexpected outcome was that neurons in the monkey's pre-frontal motor cortex fired, not only when the monkey performed the action, but also when the monkey observed a human experimenter performing the same action. As a control, it turned out that activation did not occur when the human experimenter used a tool (such as tweezers) to perform the action. The implication was that, far from being independent faculties, action and perception were somehow intimately linked.

The discovery of mirror neurons triggered an avalanche of research aimed at uncovering the implications of significant sensorimotor overlap. The basic idea was that mirror structures appeared to facilitate mental *simulations* that could be used for interpreting the actions and intentions of others [34]. Simulation not only provides a generative forward model that may be used to explain observed events, but it also facilitates the prediction of future events, the imagination of novel events and the optimal influence of future events. The mirror mechanism thus seemed to provide a basis for a number of important behaviours such as action understanding [35], imitation and learning [36], empathy and theory of mind [37] and, of most significance here, the evolution of speech and language [38], [39], [40], [41].

Since the simulation principle suggests that generative models of spoken language production could be implicated in human speech recognition and understanding, the discovery of mirror neurons sparked a revival of interest in the *motor theory of speech perception* [42]. The jury is still out as to the precise role of the speech motor system in speech perception, but see [43], [44], [45] for examples of discussion on this topic.

The mirror neuron hypothesis has also had some impact on robotics research (see [46], for example), and the notion of mental simulation as a *forward model/predictor* mechanism has inspired new theories of language [47], [48], [49] and speech perception [50] .

2.5 Perceptual Control Theory

As suggested above, the structural coupling of an agent with its environment (including other agents) could be instantiated as a one-way causal dependency. However, it is more likely that coupling would be bi-directional, and this implies the existence of a *dynamical system* with *feedback*. Feedback - in particular, *negative* feedback - provides a powerful mechanism for achieving and maintaining *stability* (static or dynamic), and feedback control systems have been posited as a fundamental property of living systems [51], [52].

Founded on principles first expounded in the field of cybernetics [53], and railing against the traditional behaviourist perspective taken by mainstream psychologists, *perceptual control theory (PCT)* focuses on the consequence of a negative-feedback control architecture in which behaviour emerges, not from an external stimulus, but from an organism's internal drive to achieve desired perceptual states [54]. Unlike the traditional stimulus-response approach, PCT is able to explain how a living organism can compensate for (unpredictable) disturbances in the environment without the need to invoke complex statistical models. For example, the orientation of a foot placed on uneven ground is controlled, not by computing the consequences of an unusual joint angle, but by the need to maintain a stable body posture. Likewise, PCT suggest that the clarity of speech production is controlled, not by computing the consequences of the amount of noise in an environment, but by the need to maintain a suitable level of perceived intelligibility.

Indeed, the importance of feedback control in speech has been appreciated for some time, coupled with the realisation that living systems have to balance the effectiveness of their actions against the (physical and neural) effort that is required to perform the actions [55]. This principle has been used to good effect in a novel form of speech synthesis that regulates its pronunciation using a crude model of the listener [56].

2.6 Intentionality, Emotion and Learning

PCT provides an insight into a model of behaviour that is active/intentional rather than passive/reactive, and this connects very nicely with the observation that human beings tend to regard other human beings, animals and even inanimate objects as *intentional* agents [57]. It also links with the view of language as an intentional behaviour [58], and thus with mirror neurons as a mechanism for inferring the communicative intentions of others [59].

Intentionality already plays a major role in the field of *agent-based modelling*, in particular using the BDI *Beliefs, Desires, Intentions* paradigm [60], [61]. BDI is an established methodology for modelling emergent behaviours from swarms of 'intelligent' agents, but it doesn't specify how to recognise/interpret behaviour under conditions of ambiguity or uncertainty. Nevertheless, BDI does capture some important features of behaviour, and it is useful to appreciate that beliefs equate to *priors* (which equate to memory), desires equate to goals, and intentions drive planning and action.

Viewing the behaviour of living systems as intentional with the consequences of any actions being monitored using perceptual feedback, leads to a model of behaviour that is driven by a comparison between desired and actual perceptual states (that is, by the error signal in a PCT-style feedback control process). This difference between intention and outcome can be regarded as an *appraisal* [62] of emotional valence whereby a match is regarded as positive/happy and a mismatch is regarded as negative/unhappy [63]. From this perspective, emotion can be seen as a driver of behaviour (rather than simply a consequence of behaviour) and provides the force behind adaptation and learning.

3 Bringing the Ideas Inside

The foregoing provides a wealth of insights from outside the technical field of spoken language processing that could have a direct bearing on future spoken language systems. In particular, it points to a novel computational architecture for spoken language processing in which the distinctions between traditionally independent system components become blurred. It would seem that speech recognition and understanding should be based on forward models of speech generation/production, and that those models should be the same as those used by the system to generate output itself. It turns out that dialogue management should be concerned less with turn taking and more with synchronising and coordinating its behaviours with its users.

The ideas above also suggest that a system's goals should be to satisfy users' rather than systems' needs, and this means that systems need to be able to model users and determine their needs by empathising with them. A failure to meet users' needs should lead to negative affect in the system, an internal variable which is not only used to drive the system's behaviour towards satisfying the user, but which could also be expressed visually or vocally in order to keep a user informed of the system's internal states and intentions. The previous section also points to a view of spoken language processing that is more integrated with its external environment, and to systems which are constantly adapting to compensate for the particular contextual circumstances that prevail.

3.1 Existing Approaches

A number of these ideas have already been discussed in the spoken language processing literature, and some are finding their way into practical systems. For example, the PRESENCE (*PREdictive SENsorimotor Control and Emulation*) architecture [64], [65], [66] draws together many of these principles into a unified framework in which the system has in mind the needs and intentions of its users, and a user has in mind the needs and intentions of the system. As well as the listening speech synthesiser mentioned earlier [67], PRESENCE has informed a number of developments in spoken language processing including the use of user emotion to drive dialogue [68], *AnTon* - an animatronic model of the human tongue and vocal tract [69] and the parsimonious management of interruptions in conversational systems [70], [71].

Another area of on-going work that fits well with some of the themes identified above is the powerful notion of *incremental* processing whereby recognition, dialogue management and synthesis all progress in parallel [72], [73], [74], [75]. These ideas fit well with contemporary approaches to dialogue management using POMDPs *Partially-Observable Markov Decision Processes* [76] [77].

However, despite these important first steps, as yet there is no mathematically grounded framework that encapsulates all of the key ideas into a practical computational architecture. Of course, this is not surprising - these are complex issues that can be difficult to interpret. So, where might one start? The following is a preliminary taste of what might be required [78].

3.2 Towards a General Conceptual Framework

One of the main messages from the foregoing is that a key driver of behaviour - including speaking - for a living system seems to be *intentionality* (based on *needs*). Consider, therefore, a *world* containing just two intentional agents - *agent*1 and *agent*2. The world itself obeys the Laws of Physics, which means that the evolution of events follows a straightforward course in which actions lead to consequences (which constitute further actions) in a continuous cycle of cause and effect.

$$Actions_t \leadsto Consequences \mapsto Actions_{t+1} \ . \tag{1}$$

The behaviour of the world can thus be characterised as ...

$$Consequences = f_{world}\left(Actions\right) \ , \tag{2}$$

where f is some function which transforms *Actions* into *Consequences*.

The two intentional agents are each capable of (i) effecting changes in the world and (ii) inferring the causes of changes in the world.

In the first case, the intentions of an agent lead to actions which in turn lead to consequences ...

$$Intentions \leadsto Actions \leadsto Consequences \ . \tag{3}$$

The behaviour of the agent can thus be characterised as ...

$$Actions = g_{agent}\left(Intentions\right) \ , \tag{4}$$

where g is some function that transforms *Intentions* into *Actions*.

In the second case, an agent attempts to infer the actions that gave rise to observed consequences.

$$Actions \leadsto Consequences \leadsto \widehat{Actions} \ . \tag{5}$$

The behaviour of the agent can thus be characterised as ...

$$\widehat{Actions} = h_{agent}\left(Consequences\right) \ , \tag{6}$$

where h is some function that transforms *Consequences* into estimated *Actions*.

This analysis becomes interesting when there is (intentional) interaction between the two agents. However, before taking that step, it is necessary to consider the interactions between the agents and the world in a little more detail.

An Agent Manipulating the World Consider an agent attempting to manipulate the world, that is intentions are transformed into actions which are transformed into consequences. In robotics, the process of converting an intention into an appropriate action is known as *action selection*, and the relevant transformation is shown in equation 4 . Note, however, the emphasis here is not on the *actions* that are required, but on the *consequences* of those actions.

$$Consequences = f_{world}\left(g_{agent}\left(Intentions\right)\right) \ , \tag{7}$$

where g is a transform from intentions to actions and, as before, f is the transform from actions to consequences.

Of course, whether the intended consequences are achieved depends on the agent having the correct transforms. It is possible to discuss how f and g might be calibrated. However, there is an alternative approach that is not dependent on knowing f or g, and that is to search over possible actions to find those that create the best match between the intentions and the observed consequences.

$$\widehat{Actions} = \underset{Actions}{\arg\min}\left(Intentions - Consequences\right) \ , \tag{8}$$

where $Intentions - Consequences$ constitutes an error signal that reflects the agent's *appraisal* of its actions. A large value means that the actions are poor; a small value means that the actions are good. Hence, the error signal can be said to be equivalent to *emotional valence* - as discussed in section 2.6. Overall, this optimisation process is a negative feedback control loop that operates to ensure that the consequences match the intentions even in the presence of unpredictable disturbances. This is exactly the type of control structure envisaged in *Perceptual Control Theory* - section 2.5.

The approach works will only function if the agent can observe the consequences of its actions. However, when an agent is manipulating another agent, the consequences are likely to be changes in internal state and thus potentially unobservable. This situation is addressed below, but first it is necessary to consider an agent interpreting what's happening in the world.

An Agent Interpreting the World The challenge facing an agent attempting to interpret what is happening in the world is to derive the actions/causes from observing their effects/consequences. If the inverse transform f^{-1} is known (from equation 2), then it is possible to compute the actions directly from the observed consequences ...

$$Actions = f^{-1}_{world}\left(Consequences\right) \ . \tag{9}$$

However, in reality the inverse transform is not known. If it can be estimated $\widehat{f^{-1}}$, then it is possible to compute an estimate of the actions ...

$$\widehat{Actions} = \widehat{f^{-1}_{world}}\left(Consequences\right) \ . \tag{10}$$

Of course the accuracy with which the causes can be estimated depends on the fidelity of the inverse transform.

An alternative approach, which aligns well with some of the ideas in the previous section, is not to use an inverse model at all, but to use a *forward model* - that is, an estimate of f $(\widehat{f})$. Estimation then proceeds by searching over possible actions to find the best match between the predicted consequences $(\widehat{Consequences})$ and the observed consequences - again, a negative-feedback control loop.

$$\widehat{Actions} = \underset{\widehat{Consequences}}{\arg\min}\left(Consequences - \widehat{Consequences}\right) \ . \tag{11}$$

Of course the forward model is itself an estimate ...

$$\widehat{Consequences} = \widehat{f_{world}}\,(Actions)\ , \tag{12}$$

which leads to ...

$$\widehat{Actions} = \underset{Actions}{\arg\min}\left(Consequences - \widehat{f_{world}}(Actions)\right)\ . \tag{13}$$

As an aside, the same idea can be expressed in a Bayesian framework, but the principle is the same - interpretation is performed using search over a forward model ...

$$\Pr(Actions|Consequences) = \frac{\Pr(Consequences|Actions)\Pr(Actions)}{\Pr(Consequences)}\ . \tag{14}$$

Hence the estimated action is that which maximises the following ...

$$\widehat{Actions} = \underset{Actions}{\arg\max}\left(\Pr(Consequences|Actions)\right)\ , \tag{15}$$

where $\Pr(Consequences|Actions)$ is the forward/generative model (equivalent to $\widehat{f_{world}}\,(Actions)$).

An Agent Communicating its Intentions to Another Agent Now it is possible to turn to the situation where one agent - *agent*1 - seeks to manipulate another agent - *agent*2. As mentioned above, in this case the consequences of *agent*1's actions may not be observable (because *agent*1's intention is to change the mental state of *agent*2). However, if *agent*1 can observe its own actions, then it can use a model to *emulate* the consequences of its actions. That is, *agent*1 uses an estimate of the forward transform $\widehat{h_{agent2}}$.

$$\widehat{Actions} = \underset{Actions}{\arg\min}\left(Intentions - \widehat{h_{agent2}}(Actions)\right)\ . \tag{16}$$

This solution is equivalent to *agent*1 actively trying out actions in order to arrive at the correct ones. However, an even better solution is for *agent*1 not to search in the real world, but to search in a *simulated* world - that is, to imagine the consequences of its actions in advance of performing the chosen ones. This is *emulation* as described in section 2.4 which introduced the action of mirror neurons.

$$\widehat{Actions} = \underset{\widetilde{Actions}}{\arg\min}\left(Intentions - \widehat{h_{agent2}}(\widetilde{Actions})\right)\ . \tag{17}$$

As before, a negative-feedback control loop manages the search and, interestingly, it can also be viewed as *synthesis-by-analysis*.

An Agent Interpreting the Actions of Another Agent For an agent to interpret the actions of another agent, they are effectively inferring the intentions of that agent. In this case, $agent2$ needs to infer the intentions of $agent1$ by comparing the observed actions with the output of a forward model for $agent1$ ...

$$\widehat{Intentions} = \underset{Intentions}{\arg\min} \left(Actions - \widehat{g_{agent1}}(Intentions)\right) . \tag{18}$$

As before, a negative-feedback control loop manages the search to find the best match and, in this case the process can be viewed as *analysis-by-synthesis.*

3.3 Using *Self* to Model *Other*

Most of the derivation thus far is relatively straightforward in that it reflects known approaches, albeit caste in an unusual framework. This is especially obvious if one maps the arrangements into a Bayesian formulation. The overlap with some of the ideas outlined in section 2 should be apparent.

However, there is one more step that serves to set this whole approach apart from more standard analyses, and that step arises from the question "where do $\widehat{g}$ and $\widehat{h}$ come from"? From the perspective of $agent1$, $\widehat{h}$ is a property of $agent2$ (and vice-versa). Likewise, From the perspective of $agent2$, $\widehat{g}$ is a property of $agent1$ (and vice-versa). The answer, drawing again on concepts emerging from section 2, is that for a particular agent - $self$ - the information required to model another agent - *other* - could be derived, not from modelling the behaviour of *other*, but from the capabilities of $self$! In other words, the simulation of *other* recruits information from the existing abilities of $self$ - just as observed in mirror neuron behaviour (section 2.4).

If spoken language is the behaviour of interest, then such an arrangement would constitute *synthesis-by-analysis-by-synthesis* for the speaker and *analysis-by-synthesis-by-analysis* for the listener.

4 Conclusion

This paper has reviewed a number of different ideas from outside the mainstream field of spoken language processing (starting from the coupling between living cells), and given an insight into how they might be integrated into a novel framework that could have some bearing on the architecture for future *intelligent interactive empathic communicative* systems [79]. The approach - which might be termed MBDIAC *Mutual Beliefs Desires Intentions Actions and Consequences* - is different from the current paradigm in that, rather than estimate model parameters off-line using vast quantities of static annotated spoken language material, it highlights an alternative developmental paradigm based on on-line interactive skill acquisition in dynamic real-world situations and environments. It also places a focus on generative models of communicative behaviour (grounded in movement and action, and generalised using metaphor) which are subsequently recruited for interpreting the communicative behaviour of others.

What is also different about the approach suggested here is that, in principle, it subsumes everything from high-level semantic and pragmatic representations down to the lowest-level sensorimotor behaviours. The approach is also neutral with respect to the sensorimotor modalities involved; hence gesture and prosody have an equal place alongside the more conventional vocal behaviours. The overall message is that it may be time to step back from worrying about the detail of contemporary spoken language systems in order to rediscover the crucial communicative context in which communicative behaviour takes place. That way we might be able to design and implement truly *Autonomous Social Agents* that are capable of conversing effectively with their human counterparts.

Finally, some readers may be tempted to think that this approach is promoting a debate between statistical and non-statistical modelling paradigms. On the contrary, the entire edifice should be capable of being caste in a probabilistic framework. The concern here is not about probability but about *priors*!

Acknowledgements

The author would like to thank colleagues in the Sheffield Speech and Hearing research group and the Bristol Robotics Laboratory for discussions relating to the content of this paper. This work was partially supported by the European Commission [grant numbers EU-FP6- 507422, EU-FP6-034434, EU-FP7-231868, FP7-ICT-2013-10-611971] and the UK Engineering and Physical Sciences Research Council [grant number EP/I013512/1].

References

1. Huang, X., Acero, A., Hon, H-W. (2001). Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR.
2. Gales, M., Young, S. (2007). The application of hidden Markov models in speech recognition. Foundations and Trends in Signal Processing, 1(3), 195–304.
3. Pieraccini, R. (2012). The Voice in the Machine. MIT Press, Cambridge, MA.
4. 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.
5. Moore, R. K. (2004). Modelling data entry rates for ASR and alternative input methods. INTERSPEECH 2004 ICSLP. Jeju, Korea.
6. Moore, R. K. (2013). Spoken language processing: Where do we go from here? In R. Trappl (Ed.), Your Virtual Butler, LNAI (Vol. 7407, pp. 111–125). Heidelberg: Springer.
7. Dawkins, R. (1991). The Blind Watchmaker. Penguin Books.
8. Gopnik, A., Meltzoff, A. N., Kuhl, P. K. (2001). The Scientist in the Crib. Perennial.
9. Moore, R. K. (2005). Towards a unified theory of spoken language processing. 4th IEEE Int. Conf. on Cognitive Informatics. Irvine, CA.
10. Maturana, H. R., Varela, F. J. (1987). The Tree of Knowledge: The Biological Roots of Human Understanding. Boston, MA: New Science Library/Shambhala Publications.

11. Garrod, S., Pickering, M. J. (2004). Why is conversation so easy? Trends in Cognitive Sciences, 8, 8–11.
12. Fusaroli, R., Raczaszek-Leonardi, J., Tyln, K. (2014). Dialog as interpersonal synergy. New Ideas in Psychology, 32, 147–157.
13. Gibson, J. J. (1977). The theory of affordances. In R. Shaw and J. Bransford (Eds.), Perceiving, Acting, and Knowing: Toward an Ecological Psychology (pp. 6782). Hillsdale, NJ: Lawrence Erlbaum.
14. Turk, M. (2014). Multimodal Interaction: A Review. Pattern Recognition Letters, 36, 189–195.
15. Wagner, P., Malisz, Z., Kopp, S. (2014). Gesture and speech in interaction: An overview. Speech Communication, 57, 209–232.
16. Mithen, S. (1996). The Prehistory of the Mind. London: Phoenix.
17. MacWhinney, B. (2005). Language evolution and human development. In D. Bjorklund & A. Pellegrini (Eds.), Origins of the Social Mind: Evolutionary Psychology and Child Development (pp. 383–410). New York: Guilford Press.
18. Tomasello, M. (2008). Origins of Human Communication. Cambridge, MA: MIT Press.
19. Clark, H. H., Brennan, S. A. (1991). Grounding in communication. Perspectives on Socially Shared Cognition. Washington DC: APA Books.
20. Pezzulo, G. (2011). Shared representations as coordination tools for interaction. Review of Philosophy and Psychology, 2, 303–333.
21. Tomasello, M. (1988). The role of joint attention in early language development. Language Sciences, 11, 6988.
22. Sebanz, N., Bekkering, H., Knoblich, G. (2006). Joint action: bodies and minds moving together. Trends in Cognitive Sciences, 10(2), 70–76.
23. Bekkering, H., de Bruijn, E. R. A., Cuijpers, R. H., Newman-Norlund, R., van Schie, H. T., Meulenbroek, R. (2009). Joint action: neurocognitive mechanisms supporting human interaction. Topics in Cognitive Science, 1, 340–352.
24. Galantucci, B., Sebanz, N. (2009). Joint action: current perspectives. Topics in Cognitive Science, 1, 255–259.
25. Steels, L. (2003). Evolving grounded communication for robots. Trends in Cognitive Science, 7(7), 308–312.
26. Roy, D., Reiter, E. (2005). Connecting language to the world. Artificial Intelligence, 167, 1–12.
27. Roy, D. (2005). Semiotic schemas: a framework for grounding language in action and perception. Artificial Intelligence, 167, 170–205.
28. Lyon, C., Nehaniv, C. L., Cangelosi, A. (2007). Emergence of Communication and Language. Springer.
29. Stramandinoli, F., Marocco, D., Cangelosi, A. (2012). The grounding of higher order concepts in action and language: A cognitive robotics model. Neural Networks, 32, 165–173.
30. Harnad, S. (1990). The symbol grounding problem. Physica D, 42, 335–346.
31. Feldman, J. A. (2008). From Molecules to Metaphor: A Neural Theory of Language. Bradford Books.
32. Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. Cognitive Brain Research, 3, 131–141.
33. Rizzolatti, G., Craighero, L. (2004). The mirror-neuron system. Annual Review of Neuroscience, 27, 169–192.
34. Wilson, M., Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. Psychological Bulletin, 131(3), 460–473.

35. Caggiano, V., Fogassi, L., Rizzolatti, G., Casile, A., Giese, M. A., Thier, P. (2012). Mirror neurons encode the subjective value of an observed action. Proceedings of the National Academy of Sciences, 109(29), 11848–11853.
36. Oztop, E., Kawato, M., Arbib, M. (2006). Mirror neurons and imitation: a computationally guided review. Neural Networks, 19, 25–271.
37. Corradini, A., Antonietti, A. (2013). Mirror neurons and their function in cognitively understood empathy. Consciousness and Cognition, 22(3), 1152–1161.
38. Rizzolatti, G., Arbib, M. A. (1998). Language within our grasp. Trends in Neuroscience, 21(5), 188–194.
39. Studdert-Kennedy, M. (2002). Mirror neurons, vocal imitation, and the evolution of particulate speech. In M. I. Stamenov & V. Gallese (Eds.), Mirror Neurons and the Evolution of Brain and Language (pp. 207–227). Philadelphia: Benjamins.
40. Arbib, M. A. (2005). From monkey-like action recognition to human language: an evolutionary framework for neurolinguists. Behavioral and Brian Sciences, 28(2), 105–124.
41. Corballis, M. C. (2010). Mirror neurons and the evolution of language. Brain and Language, 112(1), 25–35.
42. Liberman, A. M., Cooper, F. S., Harris, K. S., MacNeilage, P. J. (1963). A motor theory of speech perception. Symposium on Speech Communication Seminar. Royal Institute of Technology, Stockholm.
43. Galantucci, B., Fowler, C. A., Turvey, M. T. (2006). The motor theory of speech perception reviewed. Psychonomic Bulletin and Review, 13(3), 361–377.
44. Lotto, A. J., Hickok, G. S., Holt, L. L. (2009). Reflections on mirror neurons and speech perception. Trends in Cognitive Science, 13(3), 110–114.
45. Hickok, G. (2010). The role of mirror neurons in speech and language processing. Brain and Language: Mirror Neurons: Prospects and Problems for the Neurobiology of Language, 112(1), 1–2.
46. Barakova, E. I., Lourens, T. (2009). Mirror neuron framework yields representations for robot interaction. Neurocomputing, 72(4-6), 895–900.
47. Pickering, M. J., Garrod, S. (2007). Do people use language production to make predictions during comprehension? Trends in Cognitive Sciences, 11(3), 105–110.
48. Pickering, M. J., Garrod, S. (2013). An integrated theory of language production and comprehension. Behavioral and Brain Sciences, 36(04), 329–347.
49. Pickering, M. J., Garrod, S. (2013). Forward models and their implications for production, comprehension, and dialogue. Behavioral and Brain Sciences, 36(4), 377–392.
50. Schwartz, J. L., Basirat, A., Mnard, L., Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5), 336–354.
51. Powers, W. T. (1973). Behavior: The Control of Perception. NY: Aldine: Hawthorne.
52. Powers, W. T. (2008). Living Control Systems III: The Fact of Control. Benchmark Publications.
53. Wiener, N. (1948). Cybernetics or Control and Communication in the Animal and the Machine. New York: John Wiley & Sons Inc.
54. Bourbon, W. T., Powers, W. T. (1999). Models and their worlds. Int. J. Human-Computer Studies, 50.
55. Lindblom, B. (1990). 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.

56. Moore, R. K., Nicolao, M. (2011). Reactive speech synthesis: actively managing phonetic contrast along an H&H continuum. 17th International Congress of Phonetics Sciences (ICPhS). Hong Kong.
57. Dennett, D. (1989). The Intentional Stance. MIT Press.
58. Glock, H.-J. (2001). Intentionality and language. Language & Communication, 21(2), 105–118.
59. Frith, C. D., Lau, H. C. (2006). The problem of introspection. Consciousness and Cognition, 15, 761–764.
60. Rao, A., Georgoff, M. (1995). BDI agents: from theory to practice. Melbourne: Australian Artificial Intelligence Institute.
61. Wooldridge, M. (2000). Reasoning About Rational Agents. Cambridge, MA: The MIT Press.
62. Scherer, K. R., Schorr, A., Johnstone, T. (2001). Appraisal Processes in Emotion: Theory, Methods, Research. New York and Oxford: Oxford University Press.
63. Marsella, S., Gratch, J., Petta, P. (2010). Computational models of emotion. A Blueprint for Affective Computing-A Sourcebook and Manual, 21–46.
64. Moore, R. K. (2007). Spoken language processing: piecing together the puzzle. Speech Communication, 49(5), 418–435.
65. Moore, R. K. (2007). PRESENCE: A human-inspired architecture for speech-based human-machine interaction. IEEE Trans. Computers, 56(9), 1176–1188.
66. Moore, R. K. (2010). Cognitive approaches to spoken language technology. In F. Chen & K. Jokinen (Eds.), Speech Technology: Theory and Applications (pp. 89–103). New York Dordrecht Heidelberg London: Springer.
67. Nicolao, M., Latorre, J., Moore, R. K. (2012). C2H: A computational model of H&H-based phonetic contrast in synthetic speech. INTERSPEECH. Portland, USA.
68. Worgan, S., Moore, R. K. (2008). Enabling reinforcement learning for open dialogue systems through speech stress detection. Fourth International Workshop on Human-Computer Conversation. Bellagio, Italy.
69. Hofe, R., Moore, R. K. (2008). Towards an investigation of speech energetics using AnTon: an animatronic model of a human tongue and vocal tract. Connection Science, 20(4), 319–336.
70. Crook, N., Smith, C., Cavazza, M., Pulman, S., Moore, R. K., Boye, J. (2010). Handling user interruptions in an embodied conversational agent. AAMAS 2010: 9th International Conference on Autonomous Agents and Multiagent Systems. Toronto.
71. Crook, N. T., Field, D., Smith, C., Harding, S., Pulman, S., Cavazza, M., Charlton, D., Moore, R. K., Boye, J. (2012). Generating context-sensitive ECA responses to user barge-in interruptions. Journal on Multimodal User Interfaces, 6(1-2), 13–25.
72. Allen, J. F., Ferguson, G., Stent, A. (2001). An architecture for more realistic conversational systems. 6th International Conference on Intelligent User Interfaces.
73. Aist, G., Allen, J., Campana, E., Galescu, L., Gallo, C. A. G., Stoness, S. C., Swift, M., Tanenhaus, M. (2006). Software architectures for incremental understanding of human speech. Ninth International Conference on Spoken Language Processing: INTERSPEECH - ICSLP. Pittsburgh, PA, USA.
74. 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.
75. 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.
76. Williams, J. D., Young, S. J. (2007). Partially observable Markov decision processes for spoken dialog systems. Computer Speech and Language, 21(2), 231–422.

77. Thomson, B., Young, S. J. (2010). Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. Computer Speech and Language, 24(4), 562–588.
78. Moore, R. K. (2014). Interpreting intentional behaviour. In M. Mller, S. S. Narayanan, & B. Schuller (Eds.), Dagstuhl Seminar 13451 on Computational Audio Analysis (Vol. 3). Dagstuhl, Germany.
79. Moore, R. K. (in press). From talking and listening robots to intelligent communicative machines. In J. Markowitz (Ed.), Robots That Talk and Listen. Boston, MA: De Gruyter.