Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 418–435
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Spoken language processing: Piecing together the puzzle

Roger K. Moore *

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, South Yorkshire S1 4DP, UK

Received 23 May 2006; received in revised form 12 November 2006; accepted 30 January 2007

Abstract

Attempting to understand the fundamental mechanisms underlying spoken language processing, whether it is viewed as behaviour exhibited by human beings or as a faculty simulated by machines, is one of the greatest scientific challenges of our age. Despite tremendous achievements over the past 50 or so years, there is still a long way to go before we reach a comprehensive explanation of human spoken language behaviour and can create a technology with performance approaching or exceeding that of a human being. It is argued that progress is hampered by the fragmentation of the field across many different disciplines, coupled with a failure to create an integrated view of the fundamental mechanisms that underpin one organism's ability to communicate with another. This paper weaves together accounts from a wide variety of different disciplines concerned with the behaviour of living systems – many of them outside the normal realms of spoken language – and compiles them into a new model: PRESENCE (PREdictive SENsorimotor Control and Emulation). It is hoped that the results of this research will provide a sufficient glimpse into the future to give breath to a new generation of research into spoken language processing by mind *or* machine.

Keywords: Spoken language processing; Speech technology; Communicative behaviour; Sensorimotor control

1. Introduction

One of the greatest scientific challenges of our age is attempting to understand the fundamental mechanisms underlying spoken language processing, whether it is viewed as behaviour exhibited by human beings (Pinker, 1994; Altmann, 1997) or as a faculty simulated by machines (Holmes and Holmes, 2002). The past 50 or more years have seen tremendous progress in our appreciation of the reliability and robustness of the speech chain operating between speaker and listener, and a high degree of insight has been obtained into the principles underlying human speech perception, speech production and conversational discourse. More recently, great strides have also been made in our ability to implement an advanced spoken language technology that is capable of supporting a wide range of practical applications based on the automatic recognition and generation of speech as part of an interactive human–machine dialogue. In fact given (i) the size of the combined speech research communities (estimated to be some 10 000 individuals worldwide), (ii) the high level of research effort that has been devoted to these areas over many years and (iii) the growing visibility of spoken language processing systems in everyday life, an outsider could be forgiven for assuming that the scientific questions are just about wrapped up – all that is left is to tidy up some minor academic details.

However, as almost everyone working in these areas will readily acknowledge, the reality is that we still have a long way to go before our understanding of spoken language processing reaches a level that is capable of both providing a comprehensive explanation of human spoken language behaviour and of supporting a technology that can exhibit performance approaching or exceeding that of a human being (Lippmann, 1997; Sinha, 2002). Indeed, not only are these aspirations still far from our reach, but it is possible that simply extending our current theories and practical solutions may never lead to such a desirable state of affairs.

* Tel.: +44 11422 21807.
E-mail address: r.k.moore@dcs.shef.ac.uk

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.01.011

1.1. Bridging scientific gaps

Part of the reasoning behind this argument is that, not only are there major schisms between the different research communities addressing the issue of human versus machine spoken language processing, but knowledge is fragmented across an extremely wide range of disciplines that claim part-ownership of the area (Moore, 1993): acoustics, psycho-acoustics, phonetics, phonology, linguistics, psycholinguistics, psychology, auditory psychophysics, cognitive neuroscience, neural-imaging, human factors, signal processing, pattern recognition, computer science, machine learning, natural language processing, artificial intelligence, neuro-computing, engineering, graphics, virtual reality, interface agents, robotics etc. etc. Integrating results from all of these different areas is itself a major challenge.

Of course, this fragmentation is not unique to research in the field of spoken language processing. Since Descartes, 'scientific reductionism' has dominated as the main paradigm for understanding natural phenomena (Burke, 1995). For over 400 years, scientists have made tremendous progress across the breadth of human knowledge by making assumptions and approximations in order to partition a problem into more easily addressable sub-parts. However, the downside of the standard scientific method is that it leads inevitably to greater and greater knowledge about smaller and smaller aspects of a problem. As a result, progress towards the unification of different theories (Wang, 2003) can be slow and ponderous, and success on the scientific 'grand challenges' (Hoare and Milner, 2005) continues to elude the scientific community.

1.2. The scale of the problem

These are very important issues, not least in spoken language processing. Indeed, combining a statement by Dawkins (1991) about the complexity of human beings with an observation by Gopnik et al. (2001) concerning the sophistication of speech, it can be argued that *spoken language is the most sophisticated behaviour of the most complex organism in the known universe* (Moore, 2005a). It is therefore perhaps not surprising that, after only 50 or so years, we may still be just scratching the surface of a real and deep understanding of the fundamental mechanisms that underpin one organism's ability to communicate with another, and the special role of spoken language as a key component of cooperative and competitive social interaction between human beings.

Of course, the different scientific communities that study human spoken language and speech technology systems are not without their own ideas about where future progress might lie in their respective niche areas (Greenberg, 1996; Bourlard et al., 1996; Hermansky, 1998; Keller, 2001; Cooke, 2003; Hawkins, 2003; Lee, 2004; Morgan et al., 2005; Moore, 2005c). However, what is missing is a truly integrated view that not only draws the relevant pieces of knowledge together, but which also serves to provide a coherent explanation of what is, after all, a single behaviour.

1.3. The puzzle of spoken language

Clearly the issue being addressed in this paper is intentionally much more wide ranging than 'bridging the gap between automatic and human speech processing' (the main emphasis of this special issue). In fact the paper leads to the conclusion that this innocent and enticing phrase may in itself be entirely misleading as to the nature of the challenge facing the different research communities. However, rather than focus on the differences between the aims and achievements of the various spoken language research communities (Moore and Cutler, 2001), it may be more profitable to focus on an aspect of speech that is universally agreed to be the main scientific challenge: the immense *variability* of spoken language.

Many authors have written extensively about the inherent variation, or lack of invariance, that is manifest in speech, both in terms of its cognitive basis and linguistic expression, as well as its audio–visual realisation. For example, in her survey of 50 years of research in speech perception, Sarah Hawkins (2004) refers to the puzzle that "*that we feel we hear stable, or invariant, percepts of words and phonemes despite their enormous articulatory-acoustic variability in different contexts*", and it is precisely an attempt to capture the immense variability/unpredictability in speech that drives the speech technology community to collect larger and larger corpora of speech data with which to train their statistical models for automatic speech recognition (Everman et al., 2005) or their inventories of concatenative segments for text-to-speech synthesis (Keller, 2001).

1.4. Whither the source of variability in speech?

It could be argued that the continued prevalence of unexplained variability in speech is an indication that its source may lie *outside* of the context in which it is being studied. For example, one consequence of the fragmentation in spoken language processing research is that models of speech perception are treated somewhat *independently* from models of speech production,[1] and techniques for automatic speech recognition are developed quite *independently* from techniques for speech generation. As a result, the majority of current explanations assume a basic stimulus–response relationship between distal cues and proximal percepts (and vice versa) based on the traditional view of the speech chain as a sequence of transformations linking a speaker's production to a listener's perception (Denes and Pinson, 1973). The wider interactive and communicative function of speech tends to be sidelined, and thus any systematic behaviour (in production or perception)

[1] In the sense that the one is not actively embedded within the other.

that results from speaker–listener interaction is inevitably observed (and hence modelled) as *random* variation.

Indeed much of the recent progress in automatic speech recognition has derived from the introduction of stochastic modelling techniques specifically to handle unexplained variability within a sound mathematical (and computational) framework (Jelinek, 1998). The use of conditional probabilities allows a modest degree of prior structure to be modelled (such as phonetic context-dependency), but any behaviour which is not static or which is uncorrelated with existing model parameters is obliged to be characterised as residual unexplained variation and thus accommodated within the variances of the probability density functions. This approach is the basis of what Makhoul and Schwartz (1984) called 'ignorance-based modelling', and it has had considerable success. However, a large part of the research community appears to have forgotten that, just because the use of statistics provides the best method of modelling variability (Jelinek, 1996), it does not follow that the underlying system is not highly deterministic. This means that the search for structured models that explain systematic variation is still as important as the search for data to estimate the parameters of the models, and that the main challenge should be to reduce uncertainty in order to increase predictability.

1.5. A way forward?

This paper represents an attempt by the author to piece together the puzzle of spoken language processing. Inspiration has been drawn from a wide variety of different disciplines – many of them outside the normal realms of spoken language – and some of the latest published ideas have been combined with some older proposals that seem to have been overlooked. As one would expect, a complete solution has yet to emerge. However, the author hopes that the broad framework of connections established in this paper will provide a sufficient glimpse into the future to give breath to a new generation of research into spoken language processing by mind *or* machine.

2. Collecting the pieces

Any attempt to weave together accounts from a wide range of different disciplines that are concerned with the behaviour of living organisms in general and human beings in particular, inevitably comes up against fundamental philosophical issues such as the nature of 'intelligence', 'consciousness', 'thought' and 'emotion', as well as questions about the structure and functioning of the brain. Many of these areas are currently the subject of intense investigation, and inspiration for models of spoken language can be drawn from a number of key areas. Five common threads that seem to emerge are (i) the illusion of invariance, (ii) the power of feedback control systems, (iii) the importance of memory and imitative behaviour in predicting future events, (iv) evidence for significant overlap between sensory and motor processes and (v) the fundamental role of emotion in driving behaviour.

2.1. The illusion of invariance

The immense variability in the behaviour of living organisms has been a subject of research for a very long time, and the behavioural sciences have developed a wide variety of tools and techniques in an attempt to understand the underlying variables that condition perceptual processing and motor behaviour in both humans and other living organisms. Brunswik (1952) was the first psychologist to acknowledge the role of uncertainty in the relationship between an organism and its environment, and he established an approach known as 'probabilistic functionalism' – or the 'Brunswikian lens model' – in which proximal percepts and their distal cues are distinguished from proximal responses and their distal effects – see Fig. 1. Brunswik's model has had a significant impact on studies of human cognitive behaviour (e.g. Figueredo et al., 2006), as well as on spoken language (e.g. Scherer, 2003).

In direct contrast to the stimulus–response models typified by Brunswik, Powers (1973) criticised the traditional behavioural science view that behaviour was unpredictable and random. He observed that this perspective meant that scientists "*spent a lot of time looking for generalisations that don't depend on orderliness in behaviour* (*e.g. using stochastic approaches*)". Powers argued that behaviour was indeed orderly, not in the sense of fixed (or stochastic) patterns of stimulus–response activity, but in that it is actively shaped by an organism into repeatable states and patterns. In other words, organisms "*act so as to get what they want, in the face of unpredictable events*" (Taylor, 1999), and Powers saw this as evidence for the operation of feedback control processes. Inspired by the power of control systems

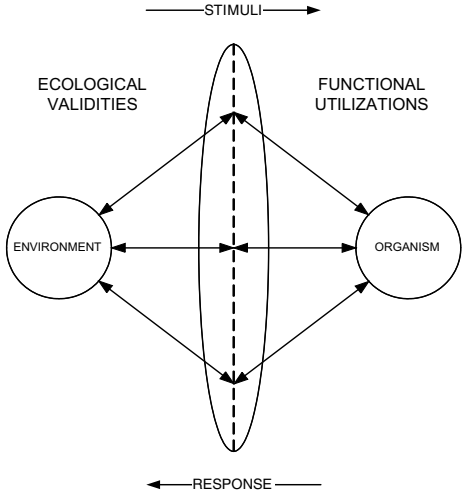

Fig. 1. The Brunswikian lens model of an organism's relationship with its environment which emphasises the one-to-many mapping between a distal event and its proximal cues (stimulus) and the many-to-one mapping between proximal means and its distal achievement (response).

to explain complex dynamical behaviour, Powers introduced the general notion of 'Perceptual Control Theory' (PCT) in which the behaviour of a living system is modelled using a hierarchy of such feedback control processes.[2] In his view, the apparent lack of invariance in behaviour was an illusion that was created as a result of ignoring the influence of feedback and from not viewing behaviour as being a consequence of perceptual control.

2.2. Perceptual control

As an example of perceptual control in action, Powers (2005) cites the ease with which a human being is able to pilot a motor vehicle in a wide variety of driving conditions simply by occasionally checking the position of the vehicle on the road and making constant adjustments to maintain the *desired* trajectory. A key property of such a feedback control process, based on a defined reference signal (i.e. "stay on the road"), is that such an architecture renders it unnecessary to make direct measurements of all the different conditions and variables that might disturb the intended direction of the vehicle (such as the speed of a side wind, the degree of camber, the angle of the bends, etc.). All that the driver needs to do is to pay sufficient attention to the *perceptual consequences* of their own behaviour and modify it accordingly. From such examples, Powers argues that behaviour is not simply a response to perceptual stimuli, but rather that *behaviour is the control of perception* (Powers, 1973).

The basic architecture of a perceptual control system is illustrated in Fig. 2. Behaviour of an organism is said to be driven by a reference signal that specifies its 'intention' (or 'needs'). Behaviour is realised through motor action in the world which may or may not have the desired 'consequences' depending on the capabilities of the organism and any disturbances that may be present. The result of the action is sensed by the organism and the perceptual 'interpretation' of the result is compared with the original intention. Any difference – as manifest in an 'error' signal – gives rise to a behavioural adjustment that is designed to bring the interpretation closer to that which was desired. The net outcome of such a negative feedback process is that behaviour is constantly modified to meet intentions in the face of varying levels and types of disturbance.

Of course the concept of a negative feedback control process is very familiar to most engineers, and the modern world could not exist without the field of Control Engineering.[3] It is therefore astounding how little impact control theory has had on the behavioural sciences, or on models of spoken language (although see Section 3.1), or indeed

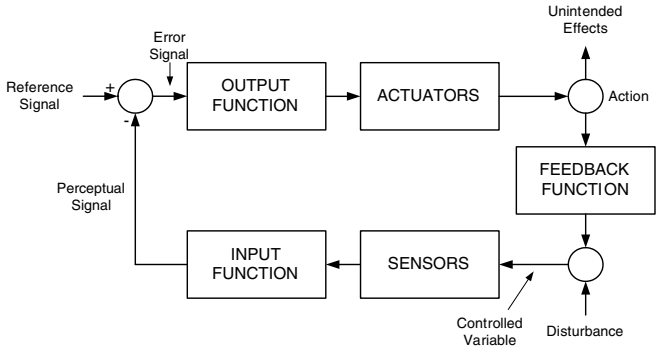

Fig. 2. Architecture of a perceptual control system.

in speech technology systems (imagine arranging to heat a room to a particular temperature by applying the currently popular machine learning paradigm of stochastic modelling[4]).

One cognitive area in which the power of closed-loop systems has been realised is in studies of the way in which human beings interact with real-time interfaces. Nicolelis (2001) observes that "*by establishing a closed loop with an artificial device, the brain … can incorporate that device … into its somatic and motor representations, and operate on them as if they were simple extensions of our bodies*".[5] This notion that closed-loop control allows an organism to view devices (or other organisms!) with which it is interacting as being components of itself may be key to understanding the mechanisms of communicative behaviour in general and spoken language in particular, as well as providing a possible source of hitherto unexplained variability.

2.3. Emulation, imitation and perceptual prediction

An interesting extension of the basic notions of PCT appears in the work by Grush (2004, 1998) on 'emulators'.

[2] Such an approach has subsequently been proposed quite independently by Grand (2003), and is posited in accounts of birdsong (Yu and Margoliash, 1996).

[3] Examples of everyday control systems include room thermostats, a cruise-control on a car, and many of the systems on board a modern aircraft.

[4] Heating an arbitrary room to a particular temperature requires the injection of just the right amount of heat based on the room's size, the presence of other sources of heat and the means for heat loss. All this can be calculated analytically, but if any of the variables change, e.g. a window is opened or more people come into the room, then these disturbances would have to be sensed, their implications measured and the overall calculations repeated. Realising that such changes are unpredictable and happening all the time, and that the number of required sensors would get out of hand, the stochastic modeller decides instead to collect a database (in an attempt to capture the unexplained variability) with which to train a probabilistic system. The resulting device gives the right temperature 95% of the time (as long as the test conditions match the training conditions) but, in order to reduce the error rate even further, the only approach that is found to work is to collect more and more data. After many years of research, there is still a residual of variability that cannot be explained, and performance asymptotes. Through all this, it has been failed to notice that a simple thermostat would have quite adequately handled the infinity of possible conditions to a defined level of accuracy.

[5] This hypothesis is related to Jeff Hawkins, (2004) observation that, from the brain's perspective, there is a lack of clarity as to where it ends and the external world begins, and to questions about how an organism distinguishes itself from others – some patients with parietal lesion cannot (Becchio et al., 2006).

Grush notes that neural feedback paths tend to be too slow to provide timely proprioceptive feedback for achieving fine motor control over fast, goal-directed movements. However, he suggests that such a limitation can be overcome using an internal model – an emulator – that generates mock versions of proprioceptive and kineasthetic feedback in response to efferent copies of the relevant motor commands. The controller gets feedback, not from the target system, but from the output of the emulator. Grush calls this 'pseudo-closed-loop control' – see Fig. 3.

Grush suggests that such an architecture not only solves the feedback timing issue but, by inhibiting the motor commands from going to the target system, it also provides a mechanism for motor imagery. From this he concludes that motor centres would be active during motor imagery (a nod towards theories evolving from the discovery of 'mirror neurons' – see Section 2.4), and that "*imagined practice should also increase motor skills*". Even more interesting, Grush goes on to hypothesise that sensory information can be processed by making a further extension to the model in which the emulator receives input from the sensory system as well as the efferent copies of the motor commands. Grush calls this a 'Kalman emulator' (after earlier work by Gerdes and Happee (1994)) because it integrates sensor information and predicted state information – see Fig. 4.

The power of the Kalman emulator architecture is that it allows perceptual filling-in. The central nervous system (controller) receives output from the emulator *not* from the sensory apparatus, and this means that the emulator's output may be much richer than its sensory input (much like the behaviour of MINERVA2 (Hintzman, 1986)). In other words, imagination (in this case in the form of a forward model) is used to complete fragmented sensory inputs.

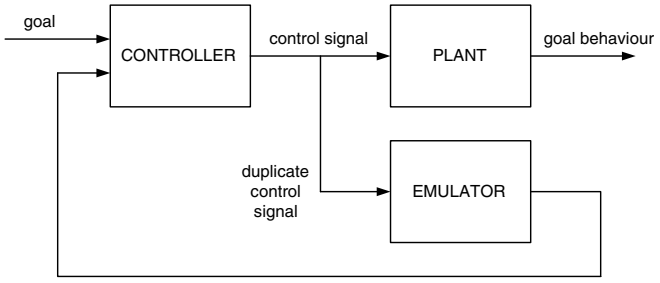

Fig. 3. Architecture of a pseudo-closed-loop control system (after Grush, 2004).

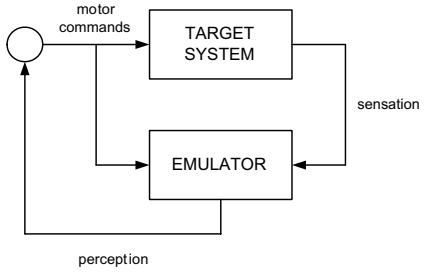

Fig. 4. Architecture of a 'Kalman emulator' (after Grush, 1998).

Similar proposals have been made by Wilson and Knoblich (2005) in order to explain the observation that the perception of another organism's behaviour activates *imitative* motor plans in the perceiver. For example, they refer to research by Fadiga et al. (2002) and others that the potentiation of muscles in a subject's own mouth increases when they listen to or watch speech. Also evidence for emulation is provided by the well-known 'Chameleon effect' (Chartrand and Bargh, 1999) in which people unconsciously mimic the behaviour of others (such as crossing their arms, or adopt similar facial expressions).[6] Wilson and Knoblich, like Grush, suggest that *covert* imitation functions as a mental simulation running in parallel to external events in order to generate top-down expectations and predictions for perception. Indeed they note that perceptual prediction is extremely common, and cite the familiar experience whereby anticipation of the next song on a favourite CD as the current one comes to an end can be so strong that you almost hear it.

It is therefore possible to hypothesise that the emulation of one's own abilities (to overcome neural transmission delays) could subsequently have been recruited in order to emulate the behaviour of others for the purposes of understanding (Cowley, 2004) and perceptual prediction (Becker, 2006). Indeed Wilson and Knoblich cite evidence that people have more knowledge of themselves than of others, and conclude that "*perceptual prediction of others is dependent on the specific qualities of one's motor programming*". This result is supported by the neuro-imaging studies of Sokhi et al. (2005) in which male subjects appeared to compare heard male voices with the internal representation of their own.

The notion of perceptual prediction is the core idea in an influential popular book by Jeff Hawkins entitled 'On Intelligence' (Jeff Hawkins, 2004). Based on Mountcastle's (1978) observation that the neocortex is remarkably uniform and hence that all areas could be performing the same basic operation, Hawkins' hypothesis is that 'intelligent' behaviour is based on what he calls a 'memory-prediction framework'.[7] Hawkins proposes that a key ingredient of intelligence is the storage of sequences in memory which are subsequently used (through a hierarchy of abstraction mechanisms) to predict what is going to happen in the external world. The purpose of the hierarchy – which is based on the six-layer columnar organisation of the cortex – is to manage the predictive framework at different levels of

[6] Interestingly, the ability to overtly imitate vocally is not universal. According to Fitch (2000), it seems that vocal mimicry is limited to human beings, birds and aquatic mammals (i.e. not apes or other primates). Jarvis (2004) hypothesises a common basis for the evolution of brain pathways for vocal learning and human language.

[7] The memory-prediction framework is being realised (and commercialised) through a technology called 'hierarchical temporal memory' (Hawkins and George, 2006).

abstraction, starting at the lowest level of patterning and only rising to higher levels if the low-level patterns are not as expected. From this Hawkins suggests that attention mechanisms would be directed by the novelty of the input, and that as unpredicted events rise in the hierarchy, so they eventually enter 'consciousness'. Hawkins places the hippocampus at the top of the neocortical pyramid.

What is interesting about Hawkins' memory-prediction framework is that it not only focuses on perceptual prediction, but also on the information that is used for prediction – in this case, episodic events stored in memory combined with derived abstractions that permit prediction by *analogy*. This is a crucial step in *generalising* from past to future experience. What is missing in the memory-prediction framework is the realisation of the intimate connection between perceptual and motor processes implied by PCT and emulation mechanisms, although Hawkins does describe a balanced system of afferent and efferent neural pathways. Clearly there is interesting potential in bringing together these different mechanisms for explaining and predicting variability, and an attempt to do so is presented in Section 5.

2.4. Mirror neurons, sensorimotor overlap and 'theory of mind'

Thus far the discussion has alluded to, but not directly addressed, the obvious conceptual links between PCT and emulation, and the relatively recent discovery of 'mirror neurons' (Rizzolatti et al., 1996; Rizzolatti and Craighero, 2004). Mirror neurons were first identified in the F5 area of premotor cortex in monkeys, where neural discharge was found to occur not only when a monkey performed an action, but also when that monkey observed a similar action being performed by another monkey. Subsequent research has confirmed the existence of such neural structures in humans, as well as other animals, and they have been implicated in the process of action understanding, intention recognition (Becchio et al., 2006), social cooperation (Pacherie and Dokic, 2006) and learning by imitation (Rizzolatti and Craighero, 2004) in conspecifics.

The notion of action understanding through access to motor planning is a direct analogue of the predictor–emulator processes discussed in the previous section. What is interesting is that there is again a strong suggestion of very close coupling between sensorimotor processes (Frith, 2002), and this is backed up by recent evidence from work on neuro-imaging (Wilson et al., 2004; Walker et al., 2004; Aboitiz et al., 2005; Warren et al., 2005). It seems that perceptual processes and motor processes in living organisms cross-refer to each other in order to support each other's prime function; motor behaviour accesses perceptual information for checking the success or otherwise of its actions, and perceptual processes access motor areas to impute underlying meaning to the actions of others.

Indeed, interpreting the behaviour of other organisms based on extrapolations from one's own behaviour appears to have close links with the general principles of 'theory of mind' (Baron-Cohen et al., 1985; Baron-Cohen, 1997), and these ideas coupled with mirror neurons have been implicated in the evolution of language (Rizzolatti and Arbib, 1998; Studdart-Kennedy, 2002; Holden, 2004). Explanations of behaviour that exclude the possibility of such sensorimotor overlap would inevitably suffer from an inability to account for key hidden dependencies, leading to an increase in the apparent level of unpredictable variation.

2.5. Emotion, affect, individuality and consciousness

The formal study of emotion in human (and animal) behaviour has a long history, from the early observational work of Charles Darwin (1872) up to the recent emergence of 'Affective Science' (Davidson et al., 2003). Over that period, three main categories of psychological model of human emotion have emerged. The earliest 'discrete' theories of emotion (stemming from Darwin's work) hypothesised the existence of a small number of basic emotions, such as happiness, sadness, fear, anger, surprise and disgust (Ekman, 1999). In such theories, it is supposed that these emotions are based on specific physiological response patterns to external stimuli. Another early model of emotion is the 'dimensional' approach (Wundt, 1874) in which a wide variety of emotions are mapped into a low-dimensional space that reflects subjective aspects of behaviour (such as positive vs. negative and active vs. passive). Douglas-Cowie et al. (2003) use such a scheme as the basis for FEELTRACE, a computer-based tool for annotating emotional data. The third, and most recent, theoretical view of emotion is the 'componential' model which emphasises the variability of different affective states, and links the production of an emotion to the *appraisal* of a situation with respect to an organism's needs and goals (Scherer et al., 2001).

The appraisal mechanism hypothesised in the componential model of emotion is clearly reminiscent of the comparison between intention and realisation within a control feedback process outlined in Section 2.2. It is possible to hypothesise that the error signal resulting from a deviation between a desired state and a perceived level of achievement represents a level of tension within a system, and thus could be viewed as a direct correlate of emotion. In a complex organism (or system) with a multiplicity of control loops, there would be a corresponding population of error signals and hence emotional states. Emotion could thus be seen as a multi-dimensional force that actually *drives* behaviour rather than simply as a response to external events (Taylor and Fragopanagos, 2005).

Emotions could not only drive behaviour, but they could also guide *attention*. Any sensory input that is perceived to be a deviation from expectations (predictions) could be treated as salient – i.e. potentially information bearing – and thus could lead to a range of behavioural adjustments such as the recruitment of additional resources, an increased weighting on appropriate sensory

channels, or an increased weighting on an error signal. In a control feedback system, the latter is equivalent to an increase in the 'loop gain', and would result in increased sensitivity (and hence, emotion).

Clearly an organism can make such adjustments 'on-the-fly' as a function of the situation it finds itself in. However, it is also possible to hypothesise that the set of default settings would, in some sense, characterise an organism's general approach to the world. Such settings could be said to constitute the 'individuality' of the organism and, depending on their values, some members of a population might be particularly sensitive, other rather slow to respond, others highly unpredictable, etc. This ties in very well with the observation by Scherer (2003) that emotion is actually a special group of behaviours within a wider set of affective states that also include mood, interpersonal stances, attitudes and personality traits.[8] Also, it is possible to hypothesise that the parameters associated with such settings could themselves be controlled by a PCT-style loop. This implies an architecture in which control systems are parasitic on others, i.e. it is not only possible to envisage a hierarchy of controls operating on various levels of intention (Powers, 1973; Grand, 2003), but also controlling the parameters of the systems carrying out the intentions.

This notion can be further extended to link up with the proposals made by Alexandrov and Sams (2005) in which they attempt to unify emotion and consciousness. Their argument, based on the fact that the mechanisms of evolution involve morphological differentiation and refinement rather than replacement, is that emotion and consciousness are essentially emergent properties of the *same* process, where there is a continuum of fine-grained emotional states between low-differentiated 'old' systems (based on behaviours such as approach and withdrawal) and highly differentiated 'new' systems. They specifically state that a "*comparison of the predicted and achieved results is the essence of consciousness*", and this ties in closely with Jeff Hawkins (2004) proposal within his memory-prediction framework that prediction failures at low levels rise up the hierarchy until they enter into consciousness. In support of their theory, Alexandrov and Sams (2005) observe that individual development goes from global 'preferenda' to detailed 'discriminanda', and that early stages of behaviour are characterised by greater emotionality.

Affective behaviour can thus be viewed, not as unpredictable variation overlaid on emotionally neutral forms, but rather as the main driving force behind all behaviour. Models that do not take this into account will be unable to access a significant conditioning variable on which much subtle behaviour might depend.

[8] It is interesting to note that excessive gain in a control feedback loop can lead to hard-limiting (extreme) behaviour, excessive delays can lead to oscillatory behaviour, and excessive damping slows response times making it difficult to react quickly enough in time-critical situations.

3. Parallels with some existing models of spoken language processing

The areas discussed above have been highlighted because they offer insights into the wider behaviour of living organisms of which spoken language can be seen to be an interesting and important special case. Whilst none of the areas have had an impact on mainstream models of human or machine spoken language processing, two – feedback control and sensorimotor overlap – have interesting parallels with some existing models, thereby lending support to the relevance of such behaviour.

3.1. Feedback control processes in spoken language

The notion that spoken language behaviour might involve feedback control processes was established by Levelt in his 'perceptual loop theory' of monitoring in speech production (Levelt, 1983, 1989, 1992, 2001; Levelt et al., 1999). Based on evidence from speech errors and repairs, but apparently unaware of Perceptual Control Theory, Levelt argued that the surprising accuracy of speech production could be explained by a process of 'self-monitoring' "*based on parsing one's own inner or overt speech*". His model – known as 'WEAVER++' – includes two feedback loops: one based on auditory feedback, and another based on the assessment of an internal pre-articulatory representation (see Fig. 5), and it has been very successful in accounting for empirical data (e.g. Hartsuiker and Kolk, 2001; Slevc and Ferreira, 2006) – thus lending support to the importance of feedback in spoken language processing (Tremblay et al., 2003).

The main emphasis of Levelt's model is on the selection of lexical items during speech production (i.e. the inner loop) rather than on the overt auditory feedback path. Of course it is well known that being able to hear your own voice has

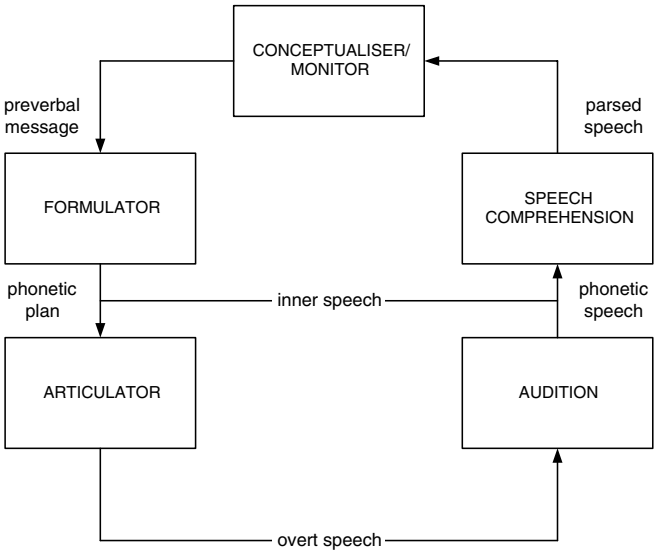

Fig. 5. Perceptual loop theory of self-monitoring in speech production (Levelt, 1983, 1989).

an effect on speaking (Bailly, 1997; Perkell et al., 1997). For example, profoundly deaf speakers can have great difficulty maintaining level control or accurate pronunciation (Geers and Moog, 1992), and delayed auditory feedback can give rise to stuttering-like behaviour (Fairbanks, 1955).

Also, it is well established that speakers alter their behaviour dynamically as a function of the communicative context. Almost 100 years ago, Lombard (1911) described how speakers in a noisy environment not only raise their level to compensate for the competing noise, but also make fine adjustments to their articulations in order to avoid any localised spectral prominences that might be present in the sound field (Lane and Tranel, 1971; Junqua, 1996). More recently, Lindblom (1990) introduced his 'H&H' (hyper-hypo) theory as a compelling explanation of the observation that "*speakers can, and typically do, tune their performance according to communicative and situational demands*". Lindblom's main argument was that the lack of invariance in speech arises because speakers constantly adjust their level of 'clarity' in order to maintain sufficient phonetic contrast. His key point was that the need to maximise discriminability is balanced by the need to minimise the energetic cost of the movements involved. As a result, he hypothesised that speakers dynamically tune their articulation between hyper- and hypo-articulation as a function of the information required for successful lexical access by the listeners. Lindblom also noted that clear speech is not simply normal speech produced louder, but that it also involves the reorganisation of articulatory gestures and acoustic patterns.

The apparent lack of invariance in speech that inspired the H&H theory is clearly highly reminiscent of the general arguments supporting PCT presented earlier. Indeed, Lindblom drew inspiration from the effectiveness of feedback control loops as the underlying mechanism for *compensatory* motor behaviour in general (i.e. not just in speech) as well as in the behaviour of other living organisms. He also pointed out that H&H contrasted directly with the mainstream stimulus–response theories of speech perception, such as the Motor Theory (Liberman and Mattingly, 1985), Quantal Theory (Stevens, 1989) and Direct Realism (Fowler, 1986).

Another area where there is dramatic evidence of PCT-style feedback-induced compensatory behaviour in spoken language is infant-directed speech or 'parentese'. Parentese is typically slower, more clearly articulated, contains exaggerated pitch contours and has a higher average pitch than adult-directed speech (Kuhl, 2004). Not only is this behaviour adopted by carers in order to be better understood, but it is also thought to play a role in encouraging imitative behaviour and thence learning by the child.

3.2. Links between speech perception and speech production

A very influential paper by Rizzolatti and Arbib (1998) describes how the human mirror neuron system includes Broca's area; hence they have proposed that mirror neurons provide a bridge between motor activity, gestural communication and the evolution of language. Studdart-Kennedy (2002) developed this line of argument in the direction of speech, proposing that speech perception and speech production must be linked if communication is to take place between speaker and listener. He observes that the Motor Theory of speech perception (Liberman and Mattingly, 1985) can be viewed as a special case of the general principle of imitative behaviour, and also proposes that Meltzoff and Moore (1997) 'active intermodal matching' (AIM) model for facial imitation could be extended to vocal mimicry.

Further neurobiological support for tight links between speech perception, speech production and speech understanding is provided by the discovery that hearing a word activates its articulatory motor programme, and understanding an action word leads to the thought of the corresponding action (Pulvermüller, 2005). Pulvermüller hypothesises that these results support existing psycholinguistic models such as the Motor Theory, but he also links the cortical basis of short-term memory with what he calls 'articulatory perception–action loops'.

4. Piecing it all together

The foregoing provides a strong but diverse base on which to build a coherent picture of intelligent behaviour in general and spoken language processing in particular. A common thread running throughout is that not only are perceptual and motor processes intimately connected through control loops that use both overt and covert sensory feedback for motor planning, but they are also linked by emulators that provide the basis for memory-based predictive behaviour that is synchronised with sensory input. Evidence for all these mechanisms operating in human spoken language processing is quite strong, and yet only a few are invoked in contemporary models or systems.

However, not all of the pieces are in place – underpinning all of the foregoing mechanisms are the fundamental factors that ultimately determine an organism's fitness to survive in an evolutionary framework: energy, time and entropy. The management of energy facilitates efficient behaviour in the context of scarce resources, the management of time facilitates efficient planning in the context of potentially harmful situations and the management of entropy facilitates efficient communications in the context of information sparsity.

4.1. Communication: entropy management

In spoken language, behaviour involves much more than executing or understanding motor activity; its primary function is active[9] *communication* between speaker and

[9] Active communication involves the voluntary or intended transfer of information from one organism to another; passive communication is involuntary or unintended, and refers to one organism's awareness of the existence of another. Both types of communication could use the same channels.

listener (Fry, 1977; Cherry, 1978). Humans, as well as many other living organisms (Brainard and Doupe, 2002; Fitch and Hauser, 2004; Meguerditchiana and Vauclair, 2006),[10] have discovered that it is possible to exploit the ability to understand the actions of others by influencing those behaviours in desirable directions (for example, to woo a mate or to warn group members of a predator). In voluntary communication, the intention is to achieve a desired effect on another organism, rather than one's own desired behaviour, and this means that the sensorimotor loop *must* include all parties.

From an evolutionary perspective, it can thus be speculated that sensory behaviour, initially established to detect the presence of food or danger was subsequently recruited to determine if one's own behaviour was achieving the desired goals. This, in turn, provided the basis of a mechanism for understanding the intentions and motivations of other organisms – especially those of conspecifics (since they are most similar to oneself, and hence most easily predicted on the basis of information drawn from one's own motivations and abilities[11]).

Language has thus evolved to exploit this ability, first by manual and vocal gesture (as evidenced by the fact that sign language and spoken language share the same neural substrate (Emmorey, 2002), and the latest results on baboon communication (Meguerditchiana and Vauclair, 2006)) then, driven by growing tension between the physical signs and the objects and events to which they refer (caused by a release from the need to *ground* the signals explicitly), by increasing abstraction towards semiotic behaviour. Once on this path – a path shared by a number of different species (Arnold and Zuberbühler, 2006) – human beings evolved an ability to handle recursive behaviour (Hauser et al., 2002; Fitch and Hauser, 2004) thereby creating a particulate structure (Abler, 1989; Studdart-Kennedy, 2002) with combinatorial properties that exploded to provide the capacity for full linguistic expression that we possess today.

It can be therefore be hypothesised that the process of evolutionary development has given rise to an increase in the entropy of the information transferred between one organism and another: from fractions of a bit per second (bps) using manual gestures to about 100 bps for human spoken language (Shannon and Weaver, 1949). One hundred bps may appear to be a very low figure in comparison to a modern digital telecommunication system (which typically transmits speech at ~13 K bps), but in reality it is extraordinarily high for a communication channel between living organisms.

[10] Interestingly, the communication systems evolved by several other species appear to exhibit the same compensatory mechanisms that are present in human speech (Doyle, 2006; Lengagne et al., 1999).

[11] From this it can be predicted that one would have increasing difficulty understanding the behaviour of organisms that are most unlike oneself, and there would be a natural tendency to anthropomorphise (even for physical objects – such as a wayward car!).

4.2. Behaviour: energy management

Another missing piece of the puzzle is the ubiquitous requirement for effective energy management in living organisms. With infinite energy resources it is very easy to plan motor behaviour – all obstacles can be overcome through sheer strength of force. For example, the quickest route between two points would always be a straight line if an organism could simply punch through anything in its path. Similarly, communication can be guaranteed if an organism is prepared to articulate at maximum clarity and maximum volume all the time. However, the reality of living systems is that energy is an extremely precious commodity, and economy of effort pervades *all* behaviour (and has done so from the dawn of evolution).

The consequence has been that energy conservation has had a strong influence on the strategies that have developed for controlling behaviour. Even the constraints that operate on the main power source for speech – the breathing mechanism and the lungs – may have a fundamental (but much overlooked) impact on the organisation and structure of spoken language (Messum, 2005). Also, Lindblom's H&H theory (see Section 3.1) explains how a pressure to minimise articulatory effort has shaped the very nature of spoken language behaviour towards a system based on relative phonetic contrast rather than absolute phonetic targets. Not only that, but an important property of the predictive nature of energy efficient perceptual processes is that resource can be allocated on the basis of the *salience* of incoming information. Communicative signals would naturally evolve to exploit the properties of such an attention mechanism, and would thus exploit un predictability subject to information theoretic constraints. Speaker behaviour then becomes one of actively managing the attentional resources of the listener for teleological goals, with both speaker and listener applying the principle of least effort to achieve their respective goals (Zipf, 1949).

In addition, selective evolutionary pressure would have favoured organisms that invoked global rather than local strategies for optimising energy usage. Successful organisms would thus inherit very effective *search* mechanisms that could be recruited for global optimisation against other kinds of criteria, and from this it is possible to see the emergence of a powerful mechanism for the selection of behaviour, i.e. planning.

4.3. Planning: time management

Living organisms are obliged to operate in real-time; all behaviour must be organised in concert with the ongoing time course of relevant events in the real world. An organism with slow reactions, or an inability to construct an appropriate solution to a problem in time, is likely to come under severe evolutionary pressure. Likewise, an organism that is obliged to find a solution by overt behaviour will not only incur a time penalty if it needs to back up, but it will also expend extra energy resources as it does so.

Simulating events using a form of internal 'virtual reality' thus not only provides an ability to discover solutions faster than real-time, but also offers the possibility of exploiting global rather than local search behaviours with considerably reduced overhead in terms of energy expenditure. Of course, if a search within such an emulation mechanism takes too long, then there would be knock-on problems for driving the real-time system. Indeed there is evidence for just such a process operating in speech production based on analysis of the behaviour of various types of disfluency (Clarke, 2002) and stuttering (Howell, 2001, 2002). In general, such catastrophic planning failures can be avoided by increasing the processing resource available to the emulation process, or by increasing the constraint on the search that it has to perform (for example by reducing the amount of available memory or by not considering all of the possibilities, i.e. reducing attention).[12]

Interestingly, the dual hypotheses that emulation involves a global search process and that emulators are invoked in understanding the behaviour of others, have direct analogues in the graph search mechanisms employed by both computational models of human word recognition (Norris, 1994; Scharenborg et al., 2003a,b, 2005) and contemporary algorithms for automatic speech recognition (Rabiner and Juang, 1993; Huang et al., 2001; Holmes and Holmes, 2002). The key difference between such graph search techniques and the emulators being proposed here is in the source of the information that is used to derive the underlying data structures. In the emulation approach such data structures are derived initially from simulations of an organism's *own* motor abilities, whereas the contemporary models of human and automatic word recognition employ models of the surface behaviour of *other* organisms.

Of course the key to planning is the ability to predict the future based on a record of the past (stored in memory). In the stochastic modelling paradigm, this is achieved through the natural abilities of probability theory to generalise through extrapolation and interpolation. However, although such an approach is attractive (especially to achieve a level of abstraction from base-level data, or to store information efficiently with the minimum of memory), it does carry the overhead of requiring substantial observational experience in order to estimate the parameters and/or structure of the models, as well as blurring the fine detail of the information that is being stored. A complimentary approach is to formulate predictive behaviour based on the ordered compilation of fragmentary traces of episodic memory (Hintzman, 1986; Goldinger, 1996, 1998; Tulving, 2002).

[12] Although the causes and explanations of stuttering are perhaps the single most contentious issue in the field of speech pathology, it is nevertheless interesting to speculate tentatively (based on the arguments in this paper) that it could arise from a lack of sufficient processing resource, from the allocation of too much memory or attention, or that the emulation and the real-time system are not sufficiently de-coupled such that covert planning behaviour leaks into the overt performance. This would suggest that the dramatic success of frequency-shifting devices in reducing stuttering (Howell, 2001) could arise from the conversion of the auditory feedback into *someone else's voice* thereby disengaging the low-level planning process and reducing the level of attentional resource allocated to speaking.

5. The PRESENCE model

It is now possible to begin to construct a model of behaviour in general and spoken language processing in particular in which speech is characterised, not in terms of individual independent static components, but as an interactive joint behaviour between participants that is conditioned on communicative context – a whole-system view in which a speaker has in mind the communicative *needs* of a listener, and a listener has in mind the communicative *intentions* of a speaker (Fujisaki, 2005) – replacing the 'speech chain' (Denes and Pinson, 1973) with the 'speech loop' (Moore, 2005b). This new model is called the PRESENCE – 'PREdictive SENsorimotor Control and Emulation' – theory of spoken language processing.

5.1. Core behaviours

The basic principle underlying PRESENCE is that it should be a sufficiently general model of behaviour that it can be applied to all but the simplest of living organisms, and thence to any artificial device that attempts to enact a behaviour normally associated with living organisms or to interact with them. In this context, and assuming that an organism is sufficiently *motivated* that it has a need to continue to exist, then the core behaviours are:

- to ***need***: an internal setting that defines a level of attainment necessary for an organism to maintain its health (e.g. Maslow (1943) hierarchy of biological, physiological, safety, belongingness, esteem and self-actualisation needs);
- to ***sense***: the ability of an organism to experience external events;
- to ***know***: a memory store containing information derived from genetic inheritance or acquired through sensory experience;
- to ***imagine***: a predicted set of events that could happen in the future (including their projected consequences) based on interpolation and extrapolation of existing knowledge using mechanisms for emulation;
- to ***intend***: a desire for a particular event to occur, for example meeting a need;
- to ***plan***: a search over all the things that could happen in order to find a sequence of events that achieve the organism's intention;
- to ***act***: selecting a behaviour in order to change the natural course of events and cause a particular event to take place;
- to ***anticipate***: a particular prediction of what might happen in the future;

- to ***perceive***: a check that anticipated events are consistent with sensory information;
- to ***attend***: the process of giving weight to sensory information which is not consistent with anticipated events, and for allocating resource in order to maximise the accuracy of prediction mechanisms;
- to ***interpret***: a search over all the things that could have happened in order to find which one fits the observed realisation and hence to decide what has happened;
- to ***feel***: a judgement of the closeness between the intention of an act and its perceived realisation;
- to ***remember***: to add experiences, interpretations and associated contextual variables to memory;
- to ***learn***: the accumulation in memory of sensorimotor experience together with the derivation of sophisticated prediction mechanisms based on similarity/analogy;
- to ***imitate***: an attempt to act out an organism's interpretation of what has happened in order to learn more about its hidden structure (for better prediction) as well as to learn how to perform it itself;
- to ***communicate***: an action that is intended to influence another organism.

These core behaviours more or less follow a logical sequence of dependencies with 'needs' as the most basic and 'communication' as the most sophisticated. However, each serves the others, and it can easily be seen how spoken language – the ultimate in communicative interaction – still plays just as much a role in everyday survival as it did in the distant evolutionary past.

5.2. Architecture

The general architecture of the PRESENCE model is illustrated in Fig. 6.

This much simplified version of PRESENCE illustrates some of the key functionality within an organism. The architecture is roughly organised into four layers. The top layer is the primary route for motor behaviour. An organism's need (need of 'self' – N_s), modulated by motivation, conditions an intention (intention of 'self' – I_s) that would satisfy that need (determined by a process of search – as indicated by the diagonal arrow running through the module), which then drives both a motor action (M_s) and an emulation of possible motor actions ($E_s(M_s)$) on the second layer. Sensory input feeds back into this second layer, and thence determines if the desired intention has been met. The large block arrows indicate that the one process is derived from the other, and the small block arrows indicate a flow of information in the opposite direction. Both are intended to represent a process of parameter sharing or 'learning'.

The third layer of the model represents a feedback path on the behaviour of 'self' based on emulating the effect of its behaviour on 'other'. In other words, $Eo(Is)$ represents the emulation by 'other' of the intentions of 'self', and $Es(Eo(Is))$ represents the emulation of that function by 'self'. A similar arrangement applies to $Es(Eo(Ms))$. The fourth layer in the model represents the organism's means for interpreting the needs, intentions and behaviour of others though a process of emulating their needs, intentions and behaviour based on their emulation of one's own needs, intentions and behaviour.

Overall, what the model attempts to capture is the general principles of the process whereby the perceived needs of others can change the needs of self, and hence give rise to totally different strategies for behaviour. The architecture illustrated is purposefully neutral with respect to the modality of an organism's interaction with the environment, or indeed the complexity of the organism involved.

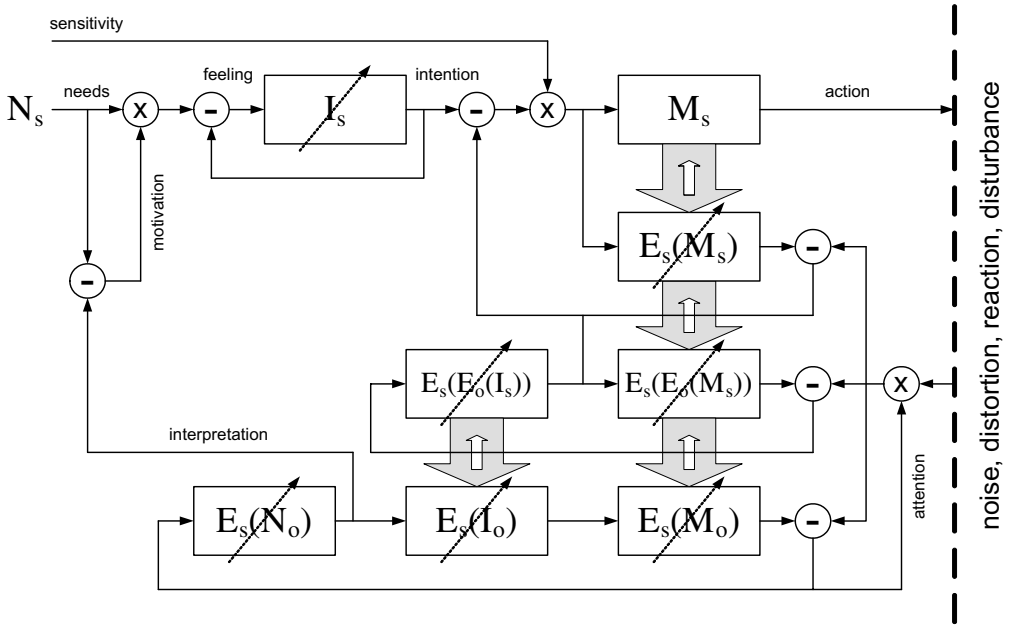

Fig. 6. Architecture of the PRESENCE model (where *S* represents 'self', *O* represents 'other', *N*: needs, *I*: intentions, *M*: motor activity, *E*( ): emulation).

A specific instantiation would require a considerably more complex arrangement, with hierarchical 'metacognitive' structures (Cox, 2005) and multiple parallel synchronous streams. However, a key feature is that the PRESENCE model effectively sidesteps the long-running bottom-up vs. top-down debate, and instead substitutes a more integrated view of sensorimotor processes.

Also, a core aspect of PRESENCE not illustrated in Fig. 6 is the related memory structures, and the processes for learning, acquisition and plasticity. Interestingly the latter has some parallels with 'adaptive critic architectures' (Barto, 1995).

6. Implications for models of spoken language processing

PRESENCE models speech as an emergent behaviour from the interaction of two (or more, depending on the number of interlocutors) parallel and integrated hierarchical perceptual control processes supporting the efficient exchange of communicative intent based on predictive emulation. It is assumed that the prime function of speaking – the communicative intent – is not to control the speaker's perception of their own voice, but to *control listener behaviour*. Therefore, all other forms of feedback are subservient to this role – even the control of a listener's perception of the linguistic message.[13] As a result, depending on the perceived success of communication, the speaker controls the level of intelligibility and comprehensibility, not simply by using more or less speaking effort, but by actively *not* saying what the listener might hear by mistake in the perceived communicative context.

Similarly, a speaker controls the listener's perception of the speaker's affective state (emotion, mood, interpersonal stances, attitudes, personality traits) and individuality. However, as with the other behaviours, this may or may not be successful – a speaker may attempt to portray themselves in a certain way, but any mismatch between actual and transmitted internal states may be detected by the listener and interpreted accordingly. This kind of adaptive behaviour can be readily seen in social situations where certain accents might have implied stereotypical associations which a speaker wishes to avoid.

What is common to all these behaviours is that they can only be controlled under *arbitrary* conditions if there is a feedback loop. So PRESENCE not only incorporates mechanisms for real-time appraisal, but also the emulation of such behaviours for assessing their putative impact prior to articulation. In this case, the speaker's emulations are based on models of the listener that the speaker has derived from the speaker's model of themselves.

From the listener's perspective, as well as inversions of the above processes for interpreting the speaker, PRESENCE also incorporates mechanisms for controlling the allocation of attentional resources such as listening effort and the weighting of sensory data. As in speaking, emulation plays a major role in interpretation; not only can information about the current state of the external world be derived before actual sensory input arrives, it can also continue quite adequately even if it does not arrive or without using additional attentional resources.

Interpretation of a speaker's behaviour within a general acoustic environment is thus seen in the PRESENCE model as a 'phase-locking' between the listener's expectations and sensory input such that attention need only be applied where expectations deviate from reality (i.e. to minimise 'listening effort'). This means that PRESENCE models perception as an active process of selective confirmation that the world is as expected.[14] This is closer to the schema-driven view of perceptual processes (Varga and Moore, 1990, 1991; Moore, 1996) than to the traditional 'bottom-up' view of a comprehensive passive analysis continuously attempting to figure out what is going on (Marr, 1982; Bregman, 1990).

6.1. Implications for human spoken language processing

Clearly it is quite straightforward to map the general attributes of the PRESENCE model to the special communicative functions of spoken language. As presented, the model does not distinguish between different representational levels as would be usual in a classic acoustic–phonetic–syntactic–semantic structure for speech. Such structures may be invoked, for example by de-composing M_s, but it may be interesting to consider the implications of viewing them as 'emergent' properties of an integrated system, rather than as explicit partitioning of the internal processes.

The key difference between PRESENCE and the standard models of human speech generation and recognition discussed earlier is the inclusion of mechanisms for perceptual prediction that facilitate the emulation of self and of others (as well as self's emulation of other's emulation of self etc!). These referential structures are intended to capture the hidden dependencies that pervade natural speech communication, and provide an explicit source of variation in speech production as well as a means to interpret such speech in the face of any communicative context.

Therefore, the new challenges that PRESENCE brings to research into human spoken language processing are mainly in the area of memory and sensorimotor overlap. What data structures are accessed during spoken word

[13] In other words, the prime goal of communication is to get a listener to do something or to tell them something for some purpose. The linguistic message may be clear, but the listener may still not act on or integrate the information until its salience (to them) is made clear (Bara, 2005).

[14] This is in accordance with theories of saccadic eye movements in visual perception (Yarbus, 1967; Slaney, 1997), 'missing data' theory (Cooke et al., 2001), Cooke's (2003) glimpsing model of speech perception, and the sampling of language that a child performs in acquiring speech (Gopnik et al., 2001). As observed by Powers (1973), a control feedback process can function quite adequately by only occasionally sampling its sensors – a process he called 'synchronous detection'.

recognition, and how are those structures related to the motor abilities of the listener? If it is acknowledged that a speaker's production is conditioned their model of a listener, what implications does that have for models of word recognition that do not invoke such assumptions? What are the consequences for mismatch between the internal models of speaker and listener, e.g. during conversation between people with different first languages?

PRESENCE would also appear to offer the possibility of unifying different levels of linguistic representation within a single explanatory framework. Can hitherto disparate areas such as prosody be similarly integrated? Is it now possible to view intonational structure within the context of a communicative loop as part of the control mechanism for directing attention at the linguistic level? Should conversational turn-taking be modelled as an emergent consequence of the interaction of two organisms with different wants and needs?

PRESENCE also offers a model on which to base explorations of language evolution and the acquisition of spoken language by children as well as second-language learners. PRESENCE points to the existence of particular configurations of data and control structures; how might these arise in evolutionary framework, is it possible to hypothesise a staged developmental process linked to anatomical structure? Is imitative behaviour an essential step towards the efficient pooling of key information resources (Chella et al., 2006), and how did the recursive particulate structure that appears to be unique to language first arise?

These, and many other questions, are stimulated by the PRESENCE model. What thus becomes clear is that, as intended, PRESENCE has the potential to draw together a wide variety of disparate areas – all within one unifying theoretical (and computational) framework – towards a comprehensive and coherent explanation of spoken language behaviour.

6.2. Implications for speech technology

The implications of PRESENCE for speech technology are potentially rather direct. For example, the PRESENCE architecture suggests a new type of speech synthesiser that would (i) listen to its own output, (ii) perceive the effect that it is having on its listeners, and (iii) modify its behaviour accordingly in order to maximise its communicative intentions in the face of situational noise and disturbance. Such a 'reactive speech synthesiser' would alter its output characteristics on-the-fly as a function of the perceived effectiveness of its intended communication, and this would be judged by the provision of a suitable (auditory and/or visual) feedback path.

Fig. 7 illustrates the architecture of an advanced text-to-speech (TTS) system in which the effectiveness of the output speech is controlled according to the perceived effect on the listener. In order to do this, it is necessary to include a model of the listener within the feedback loop – in this case an automatic speech recogniser. This means that the

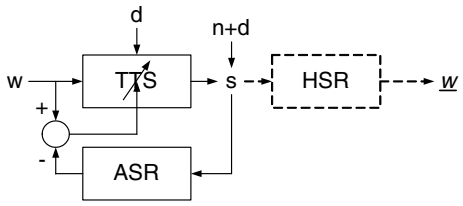

Fig. 7. PRESENCE-inspired architecture for a novel form of text (w) to speech (s) synthesiser which alters its output in order to maximise recognition accuracy in the listener in response to arbitrary noise (n) and disturbance (d).

overall system can effectively be described as 'synthesis-by-recognition' (SbR). No contemporary text-to-speech synthesiser has this capability, although something along these lines was suggested by Fallside in 1990, and Howard and Huckvale (2005) are conducting some very interesting research into training a speech synthesiser as a vocal mimic.

Of course, the architecture depicted in Fig. 7 is intended to be illustrative of the general concept. In practice, it would be necessary to invoke a rich network of control systems operating at different levels of linguistic abstraction. Such a new type of spoken language generator/synthesizer would thus be able to control and monitor its behaviour at many different layers including output volume, phonetic fidelity, choice of words and linguistic phrasing.

For perceptual interpretation, PRESENCE effectively employs a 'recognition-by-synthesis' (RbS) approach in which the emulators are *generative* models. Of course existing automatic speech recognition (ASR) systems already use generative models, usually in the form of hidden Markov models (Rabiner and Juang, 1993; Holmes and Holmes, 2002). However, an HMM is a very poor model of a speaker; it is static and lacks fine phonetic detail. PRESENCE therefore predicts a new type of speech recogniser/interpreter that, instead of HMMs, would utilize the richer structure of an actual speech generator/synthesiser based on episodic traces of actual performance. No contemporary ASR does this, although recognition-by-synthesis was first proposed by Bridle and Ralls in 1985 and some early results were published by Blomberg et al. in 1987. Fig. 8 illustrates the architecture of a system in

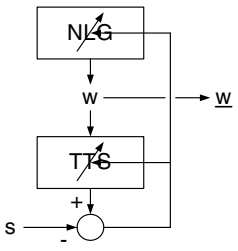

Fig. 8. PRESENCE-inspired architecture for a novel form of automatic speech recogniser that incorporates natural language and text-to-speech generators.

which speech is interpreted with respect to the output of a putative natural language and text-to-speech generator.

In practice, the architecture depicted in Fig. 8 would be expanded to include a rich network of control systems in order to reflect the complexity of structure in the putative generator. However, a particularly interesting outcome is that PRESENCE also suggests that the synthesis structures should be derived from the speech of the *listener* rather than the speaker. This rather counter-intuitive result highlights the potential benefits of establishing relationships between different sets of *speaker-dependent* models, rather than the usual approach of using speaker-independent models followed by speaker adaptation (Leggetter and Woodland, 1994).

Another compelling aspect of PRESENCE is the fact that the memory-prediction component suggests a role for episodic traces of behaviour (in both perception and production). This not only lends support to contemporary ASR research that is investigating exemplar-based representations in order to retain fine phonetic detail (De Wachter et al., 2003; Axelrod and Maison, 2004; Maier and Moore, 2005), but also has a direct analogue in contemporary unit-selection based TTS (Dutoit, 1997; Keller et al., 2001). As yet, these two areas of speech technology have not been unified into the single computational framework suggested by PRESENCE.

The simple architectures illustrated in Figs. 7 and 8 represent the first step on the road to more advanced forms of integrated automatic speech recognition and synthesis. For example, the recognition component in Fig. 7 could be substituted by the architecture in Fig. 8 (and vice versa for the synthesis components) leading to SbRbS and RbSbR. Such recursive structures are inherent in PRESENCE, and they represent a huge potential for pooling information and for parameter sharing in a practical system. The consequence is that such advanced systems would have embedded within them the means to explain the variability arising from the communicative context *without* having to be trained on ever larger quantities of speech data – truly a major step forward in the speech technology field.

Finally, although ASR and TTS are important areas of stand-alone technology, the core function of PRESENCE is to encompass the interaction between speaker and listener, in this case between a human user and a machine-based service. It will thus be necessary to incorporate research on dialogue into the PRESENCE framework, and indeed recent work in adaptive dialogue systems shows the value of employing user preference feedback and reinforcement learning to influence system behaviour (Walker et al., 2004), and this is being extended to personality (Mairesse and Walker, 2005). A more comprehensive approach would invoke a multiple interacting hierarchy of PRESENCE-based processes, each balancing individual needs and desires with an understanding of the needs and desires of a user through grounded communicative interaction in a situated and embedded environment.

7. Conclusion

The author is well aware of the dangers facing a scientist attempting to step outside the confines of their main discipline. It is very easy to appear naïve or foolish by failing to deal with the conventions and subtleties well understood by the local practitioners. Nevertheless, despite the high risks involved, this paper has attempted to draw together theoretical ideas from a wide range of different disciplines and to place them side by side in the hope that it would be possible to catch a glimpse into the wider workings of spoken language processing. It is hoped that, like a half-completed jigsaw, it will be possible to interpolate what we might expect to find where pieces are missing. In the view of the author, a coherent picture appears to be beginning to emerge in the form of the PRESENCE model. However, whether the attempt has been successful is ultimately a matter for the reader to decide and for the future to determine. Nevertheless, if these arguments hold water, then it is possible to conclude that it will *never* be possible to collect enough data to fully characterise the relationship between the linguistic message and the acoustic realisation, and that bridging the gap between human and automatic speech processing is only going to be possible if both communities step outside their usual comfort zones to consider the wider issues of human behaviour.

Acknowledgement

The author would like to thank the reviewers for their knowledgeable and constructive comments.

References

Abler, W.L., 1989. On the particulate principle of self-diversifying systems. J. Social Biol. Struct. 12, 1–13.

Aboitiz, F., Garcia, R., Brunetti, E., Bosman, C., 2005. Imitation and memory in language origins. Neural Networks 18, 1357.

Alexandrov, Y.I., Sams, M.E., 2005. Emotion and consciousness: end of a continuum. Cognitive Brain Res. 25, 387–405.

Altmann, G.T., 1997. The Ascent of Babel. Oxford University Press.

Arnold, K., Zuberbühler, K., 2006. Semantic combinations in primate cells. Nature 441, 303.

Axelrod, S., Maison, B., 2004. Combination of hidden Markov models with dynamic time warping for speech recognition. In: Proc. IEEE ICASSP.

Bailly, G., 1997. Learning to speak: sensori-motor control of speech movements. Speech Comm. 22, 251–267.

Bara, B.G., 2005. Cognitive Pragmatics. MIT Press, Cambridge, MA.

Baron-Cohen, S., 1997. Mindblindness: Essay on Autism and the Theory of Mind. The MIT Press.

Baron-Cohen, S., Leslie, A.M., Frith, U., 1985. Does the autistic child have a "theory of mind"? Cognition 21, 37–46.

Barto, A.G., 1995. Adaptive critics and the basal ganglia. In: Houk, J.C., Davis, J., Beiser, D. (Eds.), Models of Information in the Basal Ganglia. MIT Press, Cambridge, MA, pp. 215–232.

Becchio, C., Adenzato, M., Bara, B.G., 2006. How the brain understands intention: different neural circuits identify the componential features of motor and prior intentions. Conscious. Cognit. 15, 64–74.

Becker, J., 2006. Relation of neurological findings on decoupling of brain activity from limb movement to Piagetian ideas on the origin of thought. Cognitive Develop. 21, 194–198.

Blomberg, M., Carlson, R., Elenius, K., Granström, B., Hunnicutt, S., Lindell, R., Neovius, L., 1987. Speech recognition based on a text-to-speech synthesis system. In: Laver, J., Jack, M.A. (Eds.), European Conf. on Speech Technology, Edinburgh, pp. 369–372.

Bourlard, H., Hermansky, H., Morgan, N., 1996. Towards increasing speech recognition error rates. Speech Comm. 18, 205–231.

Brainard, M.S., Doupe, A.J., 2002. What songbirds teach us about learning. Nature 417, 351–358.

Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organisation of Sound Bradford Books. MIT Press, Cambridge, MA.

Bridle, J.S., Ralls, M.P., 1985. An approach to speech recognition using synthesis by rule. In: Fallside, F., Woods, W. (Eds.), Computer Speech Processing. Prentice Hall.

Brunswik, E., 1952. The conceptual framework of psychology. In: International Encyclopaedia of Unified Science, Vol. 1, University of Chicago.

Burke, J. 1995. Connections. Time Warner International.

Chartrand, T.L., Bargh, J.A., 1999. The chameleon effect: the perception-behavior link and social interaction. J. Personality Social Psychol. 76, 893–910.

Chella, A., Dindo, H., Infantino, I., 2006. A cognitive framework for imitation learning. Robot. Autonom. Systems 54, 403–408.

Cherry, C., 1978. On Human Communication: A Review a Survey and a Criticism. The MIT Press.

Clarke, H.H., 2002. Speaking in time. Speech Comm. 36, 5–13.

Cooke, M., 2003. Glimpsing speech. J. Phonetics 31, 579–584.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing data and unreliable acoustic data. Speech Comm. 34, 267–285.

Cowley, S.J., 2004. Simulating others: the basis of human cognition. Lang. Sci. 26, 273–299.

Cox, M.T., 2005. Metacognition in computation: a selected research review. Artif. Intell. 169, 104–141.

Darwin, C., 1872. The Expression of Emotions in Man and Animals. John Murray, London.

Davidson, R., Scherer, K.R., Goldsmith, H. (Eds.), 2003. Handbook of Affective Sciences. Oxford University Press.

Dawkins, R., 1991. The Blind Watchmaker. Penguin.

Denes, P.B., Pinson, E.N., 1973. The Speech Chain: The Physics and Biology of Spoken Language. Anchor Press, New York.

De Wachter, M., Demuynck, K., van Compernolle, D., Wambacq, P., 2003. Data driven example based continuous speech recognition. In: Proc. Eurospeech.

Douglas-Cowie, E., Cowie, R., Campbell, N. (Eds.), 2003. Speech and emotion. Speech Communication 40 (1–3), special issue.

Doyle, L., 2006. Talking with your mouth full: the feeding calls of the humpback whale. <http://www.space.com/searchforlife/seti_doyle_whale_060126.html>.

Dutoit, T., 1997. An Introduction to Text-to-Speech Synthesis. Kluwer Academic Publishers.

Ekman, P., 1999. Basic emotions. In: Dalgleish, T., Power, M. (Eds.), Handbook of Cognition and Emotion. John Wiley, New York, pp. 301–320).

Emmorey, K., 2002. The neural systems underlying language: insights from sign language research. In: Proc. AAAS Annual Meeting, pp. 1–4.

Everman, G., Chan, H.Y., Gales, M.J.F, Jia, B., Mrva, D, Woodland, P.C., 2005. Training LVCSR systems on thousands of hours of data. In: Proc. IEEE ICASSP, Philadelphia, pp. 209–212.

Fadiga, L., Craghero, L., Buccino, G., Rizzolatti, G., 2002. Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur. J. Neurosci. 15, 399–402.

Fairbanks, G., 1955. Selective vocal effects of delayed auditory feedback. J. Speech Hearing Disorders 4, 333–346.

Fallside, F., 1990. Synfrec: Speech synthesis from recognition using neural networks. In: Proc. ESCA Workshop on Speech Synthesis, pp. 237–240.

Figueredo, A.J., Hammond, K.R., McKierman, E.C., 2006. A Brunswikian evolutionary developmental theory of preparedness and plasticity. Intelligence 34, 211–227.

Fitch, W.T., 2000. The evolution of speech: a comparative review. Trends Cognitive Sci. 4 (7), 258–267.

Fitch, W.T., Hauser, M.D., 2004. Computational constraints on syntactic processing in a nonhuman primate. Science 303, 377–380.

Fowler, C.A., 1986. An event approach to the study of speech perception from a direct-realist perspective. J. Phonetics 14 (1), 3–28.

Frith, C., 2002. Attention to action and awareness of other minds. Conscious. Cognit. 11, 481–487.

Fry, D., 1977. Homo Loquens: Man as a Talking Animal. Cambridge University Press.

Fujisaki, H., 2005. Communication of intention and modeling the mind – lessons from a study on human–machine dialogue systems. In: Proc. Internat. Symp. on Communication Skills of Intention, Fukuoka.

Geers, A.E., Moog, J.S., 1992. Speech perception and production skills of students with impaired hearing from oral and total communication education settings. J. Speech Hearing Res. 35, 1384–1393.

Gerdes, V.G.J., Happee, R., 1994. The use of an internal representation in fast goal-directed movements: a modeling approach. Biol. Cybernet. 70, 513–524.

Goldinger, S.D., 1996. Words and voices: episodic traces in spoken word identification and recognition memory. J. Exp. Psychol.: Learn. Memory Cogn. 22 (5), 1166–1183.

Goldinger, S.D., 1998. Echoes of echoes: an episodic theory of lexical access. Psychol. Rev. 105 (2), 251–279.

Gopnik, A., Meltzoff, A.N., Kuhl, P.K., 2001. The Scientist in the Crib. Perennial.

Grand, S., 2003. Growing Up With Lucy, Phoenix.

Greenberg, S., 1996. Understanding speech understanding: towards a unified theory of speech perception. In: Proc. ESCA Workshop on Auditory Basis of Speech Perception, pp. 1–8.

Grush, R., 2004. The emulation theory of representation: motor control, imagery, and perception. Behav. Brain Sci. 27, 377–442.

Grush, R., 1998. Perception, imagery, and the sensorimotor loop. In: Esken, Heckmann (Ed.), A Consciousness Reader. Schoeningh-Verlag.

Hartsuiker, R.J., Kolk, H.H., 2001. Error monitoring in speech production: a computational test of the perceptual loop theory. Cogn. Psychol. 42, 113–157.

Hauser, M.D., Chomsky, N., Fitch, W.T., 2002. The faculty of language: what is it who, has it, and how did it evolve? Science 298, 1569–1579.

Hawkins, S., 2003. Roles and representations of systematic fine phonetic detail in speech understanding. J. Phonetics 31, 373–405.

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

Hawkins, S., 2004. Puzzles and patterns in 50 years of research on speech perception. From Sound to Sense: 50+ Years of Discoveries in Speech Communication. MIT, Cambridge, MA, USA.

Hawkins, J, George, D., 2006. Hierarchical temporal memory. Technical White Paper, Numeta Inc.

Hermansky, H., 1998. Should recognizers have ears? Speech Comm. 25, 3–27.

Hintzman, D.L., 1986. Schema-abstraction in a multiple-trace memory model. Psychol. Rev. 93, 411–427.

Hoare, T., Milner, R., 2005. Grand challenges for computing research. Computer J. 48 (1), 49–52.

Holden, C., 2004. The origin of speech. Science 303, 1316–1319.

Holmes, J., Holmes, W., 2002. Speech Recognition and Synthesis. Taylor and Francis.

Howard, I.S., Huckvale, M.A., 2005. Training a vocal tract synthesizer to imitate speech using distal supervised learning. In: Proc. SPECOM, pp. 159–162.

Howell, P., 2001. A model of timing interference to speech control in normal and altered listening conditions applied to the treatment of stuttering. In: Maassen, B., Hulstijn, W., Kent, R., Peters, KF.M., van Lieshout, P.H.M.M. (Eds.), Speech Motor Control in Normal and Disordered Speech. Uttgevrij Vantilt, Nijmegen, pp. 91–294.

Howell, P., 2002. The EXPLAN theory of fluency control applied to the treatment of stuttering by altered feedback and operant procedures. In:

Fava, E. (Ed.), Pathology and Therapy of Speech Disorders. John Benjamins, Amsterdam.
Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing. Prentice-Hall.
Jarvis, E.D., 2004. Learned birdsong and the neurobiology of human language. Ann. NY Acad. Sci. 1016, 749–777.
Jelinek, F., 1996. Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan. Speech Comm. 18, 242–246.
Jelinek, F., 1998. Statistical Methods for Speech Recognition. MIT Press.
Junqua, J-C., 1996. The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex. Speech Commun. 20, 13–22.
Keller, E., 2001. Towards greater naturalness: future directions of research in speech synthesis. In: Keller, E., Bailly, G., Monaghan, A., Terken, J., Huckvale, M. (Eds.), Improvements in Speech Synthesis. Wiley & Sons, Chichester, UK.
Keller, E., Bailly, G., Monaghan, A., Terken, J., Huckvale, M. (Eds.), . Improvements in Speech Synthesis. Wiley & Sons, Chichester, UK.
Kuhl, P.K., 2004. Early language acquisition: cracking the speech code. Nature Rev.: Neurosci. 5, 831–843.
Lane, H., Tranel, D., 1971. The Lombard sign and the role of hearing in speech. J. Speech Hearing Res. 14, 677–709.
Lee, C.-H., 2004. From knowledge-ignorant to knowledge-rich modeling: a new speech research paradigm for next generation automatic speech recognition. In: Proc. ICSLP, Korea.
Leggetter, C.J., Woodland, P., 1994. Speaker adaptation of continuous density HMMs using linear regression. In: Proc. ICSLP, pp. 451–454.
Lengagne, T., Aubin, T., Lauga, J., Jouventin, P., 1999. How do king penguins (*Aptenodytes patagonius*) apply the mathematical theory of information to communicate in windy conditions? Proc. Roy. Soc. Lond. 266, 1623–1628.
Levelt, W.J.M., 1983. Monitoring and self-repair in speech. Cognition 14, 41–104.
Levelt, W.J.M., 1989. Speaking: from Intention to Articulation. MIT Press, Cambridge, MA.
Levelt, W.J.M., 1992. The perceptual loop theory not disconfirmed: a reply to MacKay. Conscious. Cognit. 1, 226–230.
Levelt, W.J.M., 2001. Spoken word production: a theory of lexical access. Proc. Natl. Acad. Sci. 98 (23), 13464–13471.
Levelt, W.J.M., Roelofs, A., Meyer, A.S., 1999. A theory of lexical access in speech production. Behav. Brain Sci. 22, 1–75.
Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21, 1–36.
Lindblom, B., 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W., Marchal, A. (Eds.), Speech Production and Speech Modeling. Kluwer, pp. 403–439.
Lippmann, R., 1997. Speech recognition by machines and humans. Speech Comm. 22, 1–16.
Lombard, E., 1911. Le sign de l'élévation de la voix. Ann. Maladies Oreille, Larynx, Nez, Pharynx 37, 101–119.
Maier, V., Moore, R.K., 2005. An investigation into a simulation of episodic memory for automatic speech recognition. In: Proc. Inter-Speech, Lisbon, pp. 1245–1248.
Mairesse, F., Walker, M., 2005. Learning to personalize spoken generation for dialogue systems. In: Proc. EUROSPEECH'05, Lisbon, pp. 1881–1884.
Makhoul, J., Schwartz, R., 1984. Ignorance modelling. In: Perkell, J., Klatt, D.H. (Eds.), Invariance and Variability in Speech Processes. Erlbaum.
Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Freeman & Co., San Francisco.
Maslow, A.H., 1943. A theory of human motivation. Psychol. Rev. 50, 370–396.
Meguerditchiana, A., Vauclair, J., 2006. Baboons communicate with their right hand. Behav. Brain Res. 171 (1), 170–174.
Meltzoff, M., Moore, K., 1997. Explaining facial imitation: a theoretical model. Early Develop. Parenting 6, 179–192.
Messum, P., 2005. Learning to talk: a non-imitative account of the replication of phonetics by child learners. Unpublished Report, Department of Phonetics and Linguistics, University College London.
Moore, R.K., 1993. Whither a theory of speech pattern processing. In: Proc. Eurospeech, Berlin.
Moore, R.K., 1996. Critique: the potential role of speech production models in automatic speech recognition. J. Acoust. Soc. Amer. 99 (3), 1710–1713.
Moore, R.K., 2005a. Towards a unified theory of spoken language processing. In: Proc. 4th IEEE Internat. Conf. on Cognitive Informatics, Irvine, CA, USA, 8–10 August, pp. 167–172.
Moore, R.K., 2005b. Cognitive informatics: the future of spoken language processing? In: Keynote talk, SPECOM – 10th Internat. Conf. on Speech and Computer, Patras, Greece, 17–19 October.
Moore, R.K., 2005c. Research challenges in the automation of spoken language interaction. Keynote talk, COST278 and ISCA Tutorial and Research Workshop (ITRW) on Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005), Aalborg University, Denmark, 10–11 November.
Moore, R.K., Cutler, A., 2001. Constraints on theories of human vs. machine recognition of speech. In: Proc. SPRAAC Workshop on Human Speech Recognition as Pattern Classification, Max-Planck-Institute for Psycholinguistics, Nijmegen, 11–13 July, pp. 145–150.
Morgan, N., Zhu, Q., Stolcke, A., Sönmez, K., Sividas, S., Shinozaki, T., Ostendorf, M., Jain, P., Hermansky, H., Ellsi, D., Doddington, G., Chen, B., Çetin, Ö., Bourlard, H., Athineos, M., 2005. Pushing the envelope-aside. IEEE Signal Process. Mag. 22 (5), 81–88.
Mountcastle, V.B., 1978. An organizing principle for cerebral function the unit model and the distributed system. In: Edelman, G.M., Mountcastle, V.B. (Eds.), The Mindful Brain. MIT Press.
Nicolelis, M.A.L., 2001. Actions from thoughts. Nature 409, 403–407.
Norris, D.G., 1994. Shortlist: a connectionist model of continuous speech recognition. Cognition 52 (3), 163–253.
Pacherie, E., Dokic, J., 2006. From mirror neurons to joint actions. Cogn. Systems Res. 7, 101–112.
Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., Guiod, P., 1997. Speech motor control: acoustic goals, saturation effects, auditory feedback and internal models. Speech Comm. 22, 227–250.
Pinker, S., 1994. The Language Instinct. Penguin Books.
Powers, W.T., 1973. Behaviour: The Control of Perception. Aldine, Hawthorne, NY.
Powers, W.T., 2005. A brief introduction to perceptual control theory. <http:www.brainstorm-media.com/users/powers_w/whatpct.html>.
Pulvermüller, F., 2005. Brain mechanisms linking language and action. Nature Neurosci. Rev. 6, 576–582.
Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice-Hall.
Rizzolatti, G., Arbib, M.A., 1998. Language within our grasp. Trends Neurosci. 21, 188–194.
Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annu. Rev. Neurosci. 27, 169–192.
Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L., 1996. Premotor cortex and the recognition of motor actions. Cogn. Brain Res. 3, 131–141.
Scharenborg, O., ten Bosch, L., Boves, L., Norris, D., 2003a. Bridging automatic speech recognition and psycholinguistics: extending shortlist to an end-to-end model of human speech recognition. J. Acoust. Soc. Amer. 114 (6), 3023–3035.
Scharenborg, O., McQueen, J., ten Bosch, L., Norris, D., 2003b. Modelling human speech recognition using automatic speech recognition paradigms in SpeM. In: Proc. Eurospeech, Geneva, pp. 2097–2100.
Scharenborg, O., Norris, D., ten Bosch, L., McQueen, J.M., 2005. How should a speech recogniser work? Cogn. Sci. 29, 867–918.
Scherer, K.R., 2003. Vocal communication of emotion: a review of research paradigms. Speech Comm. 40, 227–256.

Scherer, K.R., Schorr, A., Johnstone, T. (Eds.), 2001. Appraisal Processes in Emotion: Theory, Methods Research. Oxford University Press, New York and Oxford.
Shannon, C.E., Weaver, W., 1949. The Mathematical Theory of Communication. University of Illinois Press.
Sinha, P., 2002. Recognizing complex patterns. Nature Neurosci. Suppl. 5, 1093–1097.
Slaney, M., 1997. A critique of pure audition. In: Rosenthal, D., Okuno, H. (Eds.), Computational Auditory Scene Analysis. Lawrence Erlbaum, Associates, pp. 27–42.
Slevc, L.R., Ferreira, V.S., 2006. Halting in single word production: a test of the perceptual loop theory of speech monitoring. J. Memory Lang. 54, 515–540.
Sokhi, D.S., Hunter, M.D., Wilkinson, I.D., Woodruff, P.W.R., 2005. Male and female voices activate distinct regions in the male brain. NeuroImage 27, 572–578.
Stevens, K.N., 1989. On the quantal nature of speech. J. Phonetics 17 (1/2), 3–45.
Studdart-Kennedy, M., 2002. Mirror neurons, vocal imitation, and the evolution of particulate speech. In: Stamenov, M.I., Gallese, V. (Eds.), Mirror Neurons and the Evolution of Brain and Language. Benjamins, Philadelphia, pp. 207–227.
Taylor, M.M., 1999. Editorial: perceptual control theory and its applications. Int. J. Human–Computer Studies 50, 433–444.
Taylor, J.G., Fragopanagos, N.F., 2005. The interaction of attention and emotion. Neural Networks 18, 353–369.
Tremblay, S., Shiller, D.M., Ostry, D.J., 2003. Somatosensory basis of speech production. Lett. Nature 423, 866–869.
Tulving, E., 2002. Episodic memory: from mind to brain. Annu. Rev. Psychol. 53, 1–25.
Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, Albuquerque, 3–6 April, pp. 845–848.
Varga, A.P., Moore, R.K., 1991. Simultaneous recognition of concurrent speech signals using hidden Markov model decomposition. In: Proc. Eurospeech, Genova, September, pp. 1175–1178.
Walker, M., Whittaker, S., Stent, A., Maloor, P., Moore, J., Johnston, M., Vasireddy, G., 2004. Generation and evaluation of user tailored responses in multimodal dialogue. Cognitive Sci. 28, 811–840.
Wang, Y., 2003. On cognitive informatics. Brain Mind 4, 151–167.
Warren, J.E., Wise, R.J.S., Warren, J.D., 2005. Sounds doable: auditory-motor transformations and the posterior temporal plane. Trends Neurosci. 28 (12), 636–643.
Wilson, M., Knoblich, G., 2005. The case for motor involvement in perceiving conspecifics. Psychol. Bull. 131 (3), 460–473.
Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nature Neurosci. 7 (7), 701–702.
Wundt, W., 1874. Grundzüge der Physiologischen Psychologie. Engelmann, Leipzig.
Yarbus, A.L., 1967. Eye Movements and Vision. Plenum Press, New York.
Yu, A.C., Margoliash, D., 1996. Temporal hierarchical control of singing in birds. Science 273, 1871–1875.
Zipf, G.K., 1949. Human Behavior and the Principal of Least Effort. Addision-Wesley Press, Cambridge.

Further Readings

Anderson, M.L., 2003. Embodied cognition; a field guide. Artif. Intell. 149, 91–130.
Anderson, M.L., Perlis, D.R., 2005. Logic, self-awareness and self-improvement: the metacognitive loop and the problem of brittleness. J. Logic Comput. 15 (1), 21–40.
Baddeley, A.D., Hitch, G.J., 1974. Working memory. In: Bower, G.A. (Ed.), Recent Advances in Learning and Motivation 8. Academic Press, New York, pp. 7–90.
Belavkin, R.V., 2004. On relation between emotion and entropy. In: Proc. AISB Symposium on Emotion, Cognition and Affective Computing, pp. 1–8.
Bryant, C.M., Jones, G.J.F., Wills, A.J., 2004. Integration of psychological models in the design of artificial creatures. In: Proc. AISB Symposium on Emotion, Cognition and Affective computing, pp. 9–20.
Deutsch, J.A., Deutsch, D., 1963. Attention: some theoretical considerations. Psychol. Rev. 70, 80–90.
Dijksterhuis, A., Meurs, T., 2006. Where creativity resides: the generative power of unconscious thought. Conscious. Cognit. 15, 135–146.
Donald, M., 1998. Mimesis and the executive suite: missing links in language evolution. In: Hurford, J.R., Studdert-Kennedy, M., Knight, C. (Eds.), Approaches to the Evolution of Language. Cambridge University Press, pp. 44–67.
Engelhardt, P.E., Bailey, K.G.D., Ferreira, F., 2006. Do speakers and listeners observe the Gricean maxim of quantity? J. Memory Lang. 54, 554–573.
Erlhagen, W., Mukovskiy, A., Bicko, E., Panin, G., Kiss, C., Knoll, A., van Schie, H., Bekkering, H., 2006. Goal-directed imitation for robots: a bio-inspired approach to action understanding and skill learning. Robot. Autonom. Systems 54, 353–360.
Feldman, J.A., 2005. On intelligence as memory. Artif. Intell. 169, 181–183.
Fenn, K.M., Nusbaum, H.C., Margoliash, D., 2003. Consolidation during sleep of perceptual learning of spoken language. Nature 425, 614–616.
Fodor, J., 2001. The Mind Doesn't Work That Way. MIT Press.
Gerken, L., Aslin, R.N., 2005. Thirty years of research in infant speech perception: the legacy of Peter Jusczyk. Lang. Learn. Develop. 1, 5–21.
de Graaf-Peters, V.B., Hadders-Algra, M., 2005. Ontogeny of the human central-nervous system: what is happening when? Early Human Develop. 82, 257–266.
Hawkins, J., 2005. Responses to reviews by Feldman, Perlis, Taylor. Artif. Intell. 169, 196–200.
Hunter, M.D., Woodruff, P.W.R., 2004. Characteristics of functional auditory hallucinations. Am. J. Psychiatry 161 (5), 923.
Hunter, M.D., Elickhoff, S.B., Miller, T.W.R., Farrow, T.F.D., Wilkinson, I.D., Woodruff, P.W.R., 2006. Proc. Natl. Acad. Sci. 103 (1), 189–194.
Ikuta, N., Sugiura, M., Sassa, Y., Watanabe, J., Akituki, Y., Iwata, K., Miura, N., Okamoto, H., Watanabe, Y., Sato, S., Horie, K., Matsue, Y., Kawashima, R., 2006. Brain activation during the course of sentence comprehension. Brain Lang. 97, 154–161.
John, E.R., 2002. The neurophysics of consciousness. Brain Res. Rev. 39, 1–28.
Junqua, J-C., Haton, J.P., 1996. Robustness in Automatic Speech Recognition. Kluwer Academic Publishers, Boston, Dordrecht, London.
Jusczyk, P.W., 1999. How infants begin to extract words from speech. Trends Cognitive Sci. 3, 323–328.
Kurzweil, R., 1990. The Age of Intelligent Machines. MIT Press.
Kurzweil, R., 1999. The Age of Spiritual Machines. Phoenix Press.
Lewicki, M.S., 2002. Efficient coding of natural sounds. Nature Neurosci. 5 (4), 356–363.
Lewis, R.L., 2000. Computational psycholinguistics. Encyclopedia of Cognitive Science. Macmillan Reference Ltd.
Liberman, A.M., Whalen, D.H., 2000. On the relation of speech to language. Trends Cognitive Sci. 4 (5), 187–196.
Martin-Loeches, M., 2005. On the uniqueness of humankind: is language working memory the final piece that made us human? J. Human Evolution 50, 226–229.
Paul, E.S., Harding, E.J., Mendl, M., 2005. Measuring emotional processes in animals: the utility of a cognitive approach. Neurosci. Biobeh. Rev. 29, 469–491.
Perlis, D., 2005. Hawkins on intelligence: fascination and frustration. Artif. Intell. 169, 184–191.
Philipson, L., 2002. Functional modules of the brain. J. Theor. Biol. 215, 109–119.

Pinker, S., 1997. How The Mind Works. Penguin Books.
Rakoczy, H., 2006. Pretend play and the development of collective intentionality. Cogn. Syst. Res. 7, 113–127.
Roy, D.K., Pentland, A.P., 1998. Learning words from natural audio–visual input. In: Proc. Internat. Conf. on Spoken Language Processing, pp. 1279–1282.
Roy, D.K., Pentland, A.P., 2002. Learning words from sights and sounds: a computational model. Cognitive Sci. 26, 113–146.
Schweizer, K., Moosbrugger, H., Goldhammer, F., 2005. The structure of the relationship between attention and intelligence. Intelligence 33, 589–611.
Searle, J., 1983. Intentionality: An Essay in the Philosophy of the Human Body. Cambridge University Press, New York.
Sternberg, S., Knoll, R.L., Monsell, S., Wright, C.E., 1988. Motor programs and hierarchical organization in the control of rapid speech. Phonetica 45, 175–197.
Sundström, P., 2005. Exploring the Affective Loop. Licenciate Thesis, Stockholm University, Stockholm, Sweden.
Taylor, M.M., 1992. Strategies for speech recognition and understanding using layered protocols. In: Laface, P., de Mori, R. (Eds.), Speech Recognition and Understanding – Recent Advances NATO ASI Series F75. Springer-Verlag, Berlin, Heidelberg.
Taylor, J.G., 2005. Jeff Hawkins and Sandra Blakeslee, on Intelligence, Times Books, 2004. Artificial Intelligence 169, 192–195.
Taylor, M.M., Farrell, P.S.E., Hollands, J.G., 1999. Perceptual control and layered protocols in interface design: II The general protocol grammar. Int. J. Human–Computer Studies 50, 521–555.
Tirassa, M., Bosco, F.M., Colle, L., 2006a. Rethinking the ontogeny of mindreading. Conscious. Cognit. 15, 197–217.
Tirassa, M., Bosco, F.M., Colle, L., 2006b. Sharedness and privateness in human early social life. Cogn. Syst. Res. 7, 128–139.
Toates, F., 2006. A model of the hierarchy of behaviour, cognition and consciousness. Conscious. Cognit. 15, 75–118.
Tummolini, L., Castelfranchi, C., 2006. From extended mind to collective mind. Cogn. Systems Res. 7, 140–150.
Wang, Y., 2006. A layered reference model of the brain (LRMB). IEEE Trans. Systems, Man, Cybernet. (Part C) 36 (2), 124–133.
Wörgötter, F., Porr, B., 2005. Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms. Neural Comput. 17, 245–319.
de Zubicaray, G.I., 2006. Cognitive neuroimaging: cognitive science out of the armchair. Brain and Cognition 60, 272–281.