PRESENCE: A Human-Inspired Architecture for Speech-Based Human-Machine Interaction

Roger K. MOORE, *Member, IEEE*

***Abstract*--Recent years have seen steady improvements in the quality and performance of speech-based human-machine interaction driven by a significant convergence in the methods and techniques employed. However, the quantity of training data required to improve state-of-the-art systems seems to be growing exponentially, and performance appears to be asymptoting to a level that may be inadequate for many real-world applications. This suggests that there may be a fundamental flaw in the underlying architecture of contemporary systems, as well as a failure to capitalize on the combinatorial properties of human spoken language. This paper addresses these issues and presents a novel architecture for speech-based human-machine interaction inspired by recent findings in the neurobiology of living systems. Called PRESENCE – 'PREdictive SENsorimotor Control and Emulation' – this new architecture blurs the distinction between the core components of a traditional spoken language dialogue system and, instead, focuses on a recursive hierarchical feedback control structure. Cooperative and communicative behavior emerges as a by-product of an architecture that is founded on a model of interaction in which the system has in mind the needs and intentions of a user, and a user has in mind the needs and intentions of the system.**

***Index Terms*—automatic speech recognition, speech synthesis, spoken language dialogue.**

I. INTRODUCTION

THERE are many compelling arguments to support the continued development of speech-based human-machine interaction. The majority of protagonists cite the inherent 'naturalness' of speech-enabled interfaces in which the spoken language skills acquired by users as infants can be readily recruited to understand the information provided by the output of a text-to-speech synthesizer, to control equipment by speaking to an automatic speech recognizer, or to access information by conversing with a spoken language dialogue system [1]. Even those who question the naturalness of such interactions nevertheless concede that the speech channel has the potential to offer genuine application benefits in hands-free eyes-free operational environments, where even an errorful speech-based human-machine interface can support higher rates of information transfer than competing interface technologies [2].

Manuscript received September 12, 2006.

R. K. Moore is Professor of Spoken Language Processing in the Computer Science Dept. at the University of Sheffield, S1 4DP UK (telephone: +44-114-222-1807, e-mail: r.k.moore@dcs.shef.ac.uk).

However, recent years have seen a significant convergence in the methods and techniques employed to develop speech-based human-machine interaction, and the data-driven statistical-modeling paradigm (such as hidden Markov model based acoustic modeling, n-gram based language modeling, and concatenative speech synthesis) has come to dominate the research agenda. Of course, this convergence of modeling paradigms has come about because of the very real improvements in system quality and performance that such approaches have provided over a period of almost three decades. The principle of defining a model, estimating its parameters from example data, and then deploying that model as a mechanism for generalizing in novel situations is above reproach, and the use of statistical methods represents one of the most powerful and effective tools that the scientific community has at its disposal for performing such modeling [3]. The only problem is that the quantity of training data required to improve state-of-the-art speech-based systems seems to be growing exponentially (despite the relatively low complexity of the underlying models), and system performance appears to be asymptoting to a level that may be inadequate for many real-world applications [4], [5]. Also, current speech technology is quite fragile, even in fairly benign everyday conditions; not only is contemporary automatic speech recognition quite poor at recognizing and understanding heavily accented or conversational speech, but machine generated speech lacks individuality, expression and communicative intent, and spoken language dialogue systems are rigid and inflexible.

These shortfalls in the capabilities of automated spoken language systems are in direct contrast to human spoken language behavior which is exceptionally robust and flexible - characteristics that allow human conversation to function very reliably even in difficult or extreme real-world conditions. For example, human sentence recognition accuracy is near perfect at -3dB signal-to-noise-ratio, and human generated speech is highly expressive and communicative. Such

differences between human and machine behavior suggest that there may be a fundamental flaw in the underlying architecture of contemporary systems for speech-based human-machine interaction.

This paper addresses these issues and presents a novel architecture for future speech-based human-machine interaction based on the 'PREdictive SENsorimotor Control & Emulation' (PRESENCE) theory of spoken language processing introduced by the author in [6]. A unique feature of PRESENCE is that it has been founded on results from a range of neurobiological scientific disciplines outside the normal realms of speech and language; disciplines that are aimed at understanding and modeling the communicative behaviors of living systems in general, as well as addressing the special cognitive abilities of human beings. This paper extends these results to encompass speech-based human-machine interaction, and discusses the architectural implications for future speech-enabled systems.

II. Insights from Living Systems

During the 1970s numerous attempts were made to invoke knowledge about the structure and behavior of human spoken language in order to develop practical systems for human-machine interaction. This was the era of the 'speech understanding system' [7], and it was assumed that the classical principles of phonetics and linguistics could be used to improve impoverished technological approaches. The practical outcomes were almost universally disappointing [8] with the best system using the least amount of phonetic and linguistic knowledge [9]. The perceived value of any insight into the human process has been greatly diminished ever since.

Apart from the cultural mismatch between the different research communities concerned, the difficulties encountered arose from the technologists' failure to grasp the importance of the *communicative* nature of speech coupled with the speech scientists' naïve understanding of

computational mechanisms. Both communities have subsequently retreated into their own domains and, apart from a few notable exceptions [10]-[19], very little research has attempted to 'bridge the gap'. Indeed it is interesting to note that the technology for text-to-speech synthesis (TTS) has now evolved almost as far away from models of human speech production as it is possible to be; the early approaches based on human-inspired articulatory modeling and formant synthesis have now almost completely given way to contatenative unit-selection approaches that appear to have very little analogy with the structure and behavior of the human vocal apparatus.

Of such attempts to bridge the gap between automatic speech recognition (ASR) and human speech recognition (HSR), one common approach is to modify the front-end signal processing of an ASR system to more closely reflect the characteristics of the human auditory system [10] and/or to detect linguistically-motivated features in the incoming signal [11]-[12]. Another approach has been to attempt to break down the core modeling assumptions (stationarity and 1st-order temporal dependencies) embedded within the conventional HMM paradigm by invoking a segmental structure that should be better able to characterize the coarticulatory dependencies and phonological constraints observed in everyday speech signals [14], [15]. Of particular interest is the recent work of Scharenborg et al [16] in which 'SHORTLIST' (the most highly regarded and widely accepted psycholinguistic model of human word recognition [17]) has been interfaced directly to a conventional ASR front-end, thereby creating 'SPeM' the first end-to-end HSR model (SHORTLIST assumes a phonetic transcription as input, whereas SPeM uses actual speech). Other approaches involve simulating models of human memory in order to retain the fine phonetic detail embedded in episodic traces of input speech (rather than blurring such detail within a statistical modeling framework) [18] and investigating the possibility of training ASR systems on the exaggerated characteristics of child-directed speech rather than on the reduced

forms typical of adult speech [19].

All such attempts to bridge the gap show some promise in terms of achieving comparable performance with that attainable using a conventional approach, but none seem to offer the order-of-magnitude jump in capability that is needed to match human behavior [20]. Indeed, although SPEM successfully captures many of the behaviors of HSR in an end-to-end model, and has thus attracted a significant amount of attention in both communities, its ability to recognize speech is actually lower than a conventional ASR system!

As a consequence of this situation, it is the opinion of the author that the challenge facing *both* the speech science and speech technology communities is no longer one of how to share disjoint views of a subject of common interest (viz. spoken language). Rather, the issue now appears to be how *both* communities can assimilate research results from the many disciplines outside of speech and language that are making significant progress in modeling and understanding the complex behavior of living systems in general, and the cognitive abilities of human beings in particular. For example, recent years have seen significant advances being made in the fields of neurobiology and cognitive neuroscience, and a number of these areas are delivering dramatically new insights into intelligent behavior - insights that may have a direct bearing on future models of spoken language interaction. In particular, the author has identified four areas of research that may have significant implications for the future architecture of speech-based human machine interaction. These are; (i) the growing evidence for an intimate relationship between sensor and motor behavior in living organisms, (ii) the power of negative feedback control to accommodate unpredictable disturbances in real-world environments, (iii) mechanisms for imitation and mental imagery for learning and modeling, and (iv) hierarchical models of temporal memory for predicting future behavior and anticipating the outcome of events.

A. Sensorimotor overlap

The author has argued in [6] that a key failure in both the speech technology and the speech science communities has been the natural tendency to decompose spoken language processing into its apparently obvious component parts - speech recognition, speech generation, and spoken language dialogue - and to conduct research in each area more or less *independently*. As a consequence, this enforced separation of perception, production, and interaction has made it virtually impossible for any of these fields to capitalize on theories of human behavior that hypothesize a more intimate relationship between sensor and motor activity.

Outside the narrow confines of speech research there has been considerable excitement in the field of neurocognition as a result of the discovery in the 1990s of 'mirror neurons' in the pre-motor cortex of macaque monkeys [21], [22] – ensembles of neurons that are activated, not only during a specific motor activity (such as grasping), but also during the observation of that same activity when performed by another individual. The implication is that motor planning behavior plays a key role in perceptual processes, and that the actions of others are interpreted with respect to an organism's own abilities to execute the observed behavior [23].

The discovery of mirror neurons, and the confirmation of the existence of such structures in the human brain, is having a huge impact on models of action-understanding. Indeed it turns out that an overlap of sensorimotor processing is implicated in a wide range of *intelligent* behaviors. For example, mirror neurons have been cited as a possible explanatory mechanism for models of emotion [24], consciousness [25], and of particular interest here, speech and language [26], [27]. Indeed, Rizzolatti and Arbib [26] and others [28] argue that the emergence of an audio-visual mirror system in the F5 area of the frontal cortex in close proximity to Broca's area provides a credible explanation of how spoken language evolved from more primitive communication

systems based on manual gestures.

Sensorimotor overlap, therefore, appears to be an essential architectural design that underpins the behavior of intelligent living systems, and is thought to have played a central role in the emergence of speech-based human-*human* interaction. As yet, no practical architecture for speech-based human-*machine* interaction exploits the parameter-sharing opportunities offered by creating an intimate relationship between speech input and speech output. However, the evidence for brain mechanisms linking language and action [29], and the discovery that speech sounds have been shown to activate the articulatory system [30], [31] are enticing (and are even reviving interest in Liberman's early 'motor theory' of speech perception [32]).

B. Perceptual control

Another consequence of decomposing spoken language processing into its component parts is that this reductionist approach leads to the situation where any systematic variation in behavior that arises from speaker-listener interaction is obliged to be observed (and hence modeled) as unpredictable and random. An alternative is to pursue a whole-system approach in which spoken language is modeled, not as a 'chain' of transformations from the mind of the speaker to the mind of the listener [33], but as an emergent behavior of a complex layered *control feedback system* in which a speaker has in mind the *needs* of a listener and a listener has in mind the *intentions* of a speaker.

In fact, such an approach - known as 'perceptual control theory' (PCT) [34] - was introduced in the early 1970s as a means for modeling a wide range of human cognitive behavior (including spoken language). Unfortunately, PCT is not well known outside its small group of enthusiasts, and it has been mainly directed towards explaining social and psychological phenomena.

The basic idea in perceptual control theory is that much of the apparent random variability in

human behavior can be explained using a hierarchy of closed-loop negative feedback systems. A controlling ‘reference’ variable sets the desired value of a controlled ‘output’ variable. The latter is sensed and its value is compared with the reference. The resulting error signal then drives the system in a direction that minimizes the difference.

The advantage of a negative feedback closed-loop control system is that it is capable of maintaining a controlled variable at a prescribed value in the face of an infinite number of possible disturbances. For example, a simple room thermostat maintains a constant temperature despite the opening and closing of doors and windows and changes in the external weather. The alternative – an ‘open-loop’ control system – would require a multitude of sensor arrays (e.g. to detect the degree of opening of each aperture) and a complex analytical model to calculate the required input.

PCT claims that the behavior of a living organism is actively directed towards maintaining desired perceptual consequences of its actions. This approach has the rather radical outcome that perception is viewed, not as a process for a detailed analysis of the world (including the behavior of other organisms) in order to figure out from first principles what is taking place, but as a process for checking that the world is as an organism wants. If the world is not as desired then (motor) action can be taken to make it so.

The structure of a basic perceptual control system is shown in Fig. 1. An organism’s ‘intention’ or ‘need’ is realized as an action, the consequences are sensed, and the ‘interpretation’ of the consequences is compared with the original intention. If the perceived consequences do not match the original intention, further action is automatically triggered to correct the difference. PCT thus provides a mechanism whereby the behavior of a living system is actively controlled in order to meet internal *needs*, and that the success or failure of any

particular action is judged by comparing the desired intentions against the perceived achievements. Behavior is then altered such as to achieve the desired internal state. The consequence is that an organism can easily and efficiently compensate for the infinity of potential disturbances that pervade real-world environments and obstruct it from achieving its intended goals. The apparent random variation in behavior is thus seen to be simply the external manifestation of such compensatory activity.

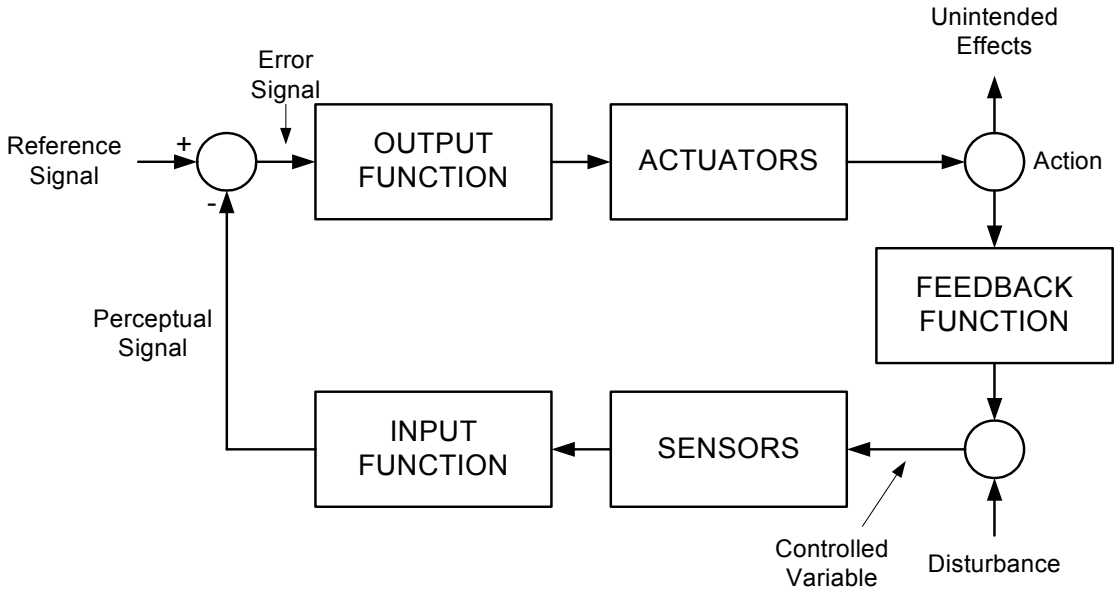

Fig. 1. Illustration of a basic perceptual control architecture.

The structure illustrated in Fig. 1 is a single layered system. However, PCT proposes a hierarchy control in which each layer defines the reference signal for the layer below. In this way, PCT is able to capture everything from low-level motor control to high-level psychological and social behavior. The levels originally hypothesized by Powers are; 1st-order: intensity; 2nd-order: sensation/vector; 3rd-order: configuration; 4th-order: transitions; 5th-order: sequence; 6th-order: relationships; 7th-order: program; 8th-order: principles; 9th-order: system concepts [34].

PCT thus provides a powerful explanatory mechanism for the complex behavior of living systems interacting with their changing physical environment by means of a hierarchy of

feedback control processes. However, also implicit in PCT is the dependency of one organism's behavior on another's. PCT thus provides the foundation for a multi-layered model of interaction between different organisms, as well as between humans and machines [35].

There is considerable evidence for the existence of perceptual control operating in spoken language. For example, it is well known that being able to hear your own voice has an effect on speaking: profoundly deaf individuals can have great difficulty maintaining clear pronunciations or achieving an appropriate level of loudness, delayed auditory feedback can cause stuttering-like behavior, and individuals naturally tend to speak louder/differently in noise [36].

Also, there is evidence that speakers actively control their spoken language behavior as a function of their listener. For example, speakers constantly adjust the fidelity of their pronunciation in order to maintain an efficient balance between communicative effectiveness and articulatory effort [37]. There is also the well known phenomenon of 'parentese' in which carers naturally exhibit quite extreme prosodic and phonetic behavior in order to be better understood by very young children [38].

Therefore, possible control variables in spoken language generation include (i) listener behavior, (ii) a listener's perception of the linguistic message, (iii) the speaker's affective state, and (iv) the speaker's individuality. Similarly, possible control variables in spoken language interpretation include (i) the listener's attention, i.e. the allocation of (computational) resources and the weighting of sensory input data, and (ii) the listener's expectations, i.e. predictions of a speaker's behavior. Interestingly, the latter can be viewed as a generative model (i.e. a model of the speaker), and this gives a clue as to the linking between perception and production that is implicit in perceptual control theory. What is important is that all these factors can only be controlled under different conditions if there is a control feedback loop.

C. Emulation of self and of others

The power of negative feedback control mechanisms for modeling complex behavior in living systems has also been investigated by scientists completely independently of PCT (e.g. [39] and [40]). One of the issues addressed by such research has been how an organism maintains accurate motor control under real-time constraints despite there being significant neurological loop delays between motor activity and proprioceptive sensor feedback. The solution in living systems is thought to be based on the evolution of mechanisms that *emulate* the effects of the intended motor actions in which the necessary feedback is provided, not by sensing the real-world, but by observing the output of the emulator in a 'pseudo' closed-loop architecture – see Fig. 2. As a result, it is possible to achieve much more rapid control than would be possible with the direct (but delayed) proprioceptive feedback path.

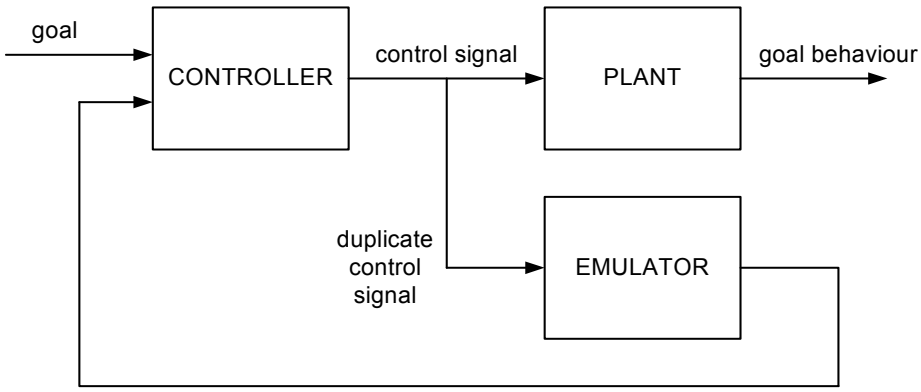

Fig. 2. Illustration of a 'pseudo' closed-loop control architecture.

The process of emulation provides an organism with the ability to simulate part (or all) of its own behavior – a form of 'mental imagery'. However, not only can this mechanism be used to overcome practical constraints such as loop delay, but it could also provide a means for advanced *planning* behavior; i.e. the ability to implement a faster than real-time simulation of possible future actions would facilitate the discovery and selection of optimal behavioral

strategies. Furthermore, it has been hypothesized that a mechanism for performing simulations of self could also be recruited to model the behavior of others [41], either as an explanatory tool for interpreting other's observed behavior (c.f. mirror neurons) or as a predictive tool for anticipating other's future behavior. These are critical behaviors for intelligent organisms, and the underlying mechanisms are clearly highly relevant to human-machine interaction.

Several authors have also noted that such emulation mechanisms would provide a foundation for imitation and *mimicry*, and hence learning [42] – processes that are thought to be highly relevant to the evolution of spoken language as a communicative behavior [43], and for the acquisition of spoken language by infants [44].

D. Hierarchical temporal memory

Another potentially relevant scientific advance outside the field of spoken language is Hawkins' 'memory-prediction framework' [45] – a relatively recent set of proposals aimed at explaining the function and purpose of the mammalian neo-cortex. Based on Mountcastle's observation that the structure of the neo-cortex is surprisingly uniform [46], Hawkins argues that its primary function is the *prediction* of future events based on past events stored in memory, and that prediction is the basis for 'intelligent' behavior in a living system. Hawkins proposes that prediction is achieved through a process of extrapolation and generalization over a hierarchy of abstract levels derived from and stored in memory [47], and he links this to the six-layered structure of the physical cortex. Such a structure is termed 'hierarchical temporal memory' (HTM).

The basic idea is that an HTM attempts to infer the *causes* of the input patterns that it receives. For example, what is the ultimate cause of the pattern of sound entering the auditory system? It is assumed that the lowest level representations are organized topologically, and that information

flows through the system over a period of time during which the external cause is assumed to be relatively static. Learning then involves the development of probabilistic internal representations - *beliefs* - from the incoming spatio-temporal patterns, starting with simple low-level causes and then moving on to more complex high-level structures. From such hierarchical representations stored in memory, it is proposed that it should be possible to construct predictions of future events in order to (i) overcome ambiguity arising from noisy or missing data, (ii) facilitate the invention of novel situations, and (iii) direct motor behavior.

HTM relates to other research linking temporal sequence modeling with neurological structure [48] as well as neurologically-inspired reinforcement learning techniques such as Barto's 'actor-critic' architecture [49] in which feedback in a control system is provided by a component – the 'critic' – that assesses both internal and external performance – a behavior thought be a property of the basal ganglia. The detailed mechanisms underlying HTM are still in their infancy, and physical implementations have yet to demonstrate advanced behavior on standard pattern processing tasks. Nevertheless, the general principles espoused in Hawkins' memory-prediction framework offer a fresh insight into the role of memory in intelligent systems, and clearly provide a candidate mechanism for the emulation capabilities discussed above as well as presenting interesting challenges to the predictive modeling paradigms employed by contemporary spoken language systems.

III. PRESENCE: AN ARCHITECTURE FOR FUTURE SPEECH-BASED HMI?

The foregoing section has summarized four key developments in modeling and understanding the neurobiological behavior of living systems, and it is clear that not only is there considerable compatibility between the different explanatory principles, but there is also a very high degree of relevance to speech-based human-human and human-machine interaction. An initial attempt to

draw these different threads together has been presented by the author in [6], and a preliminary proposal has been made for a unified theory of intelligent communicative behavior termed PRESENCE – 'PREdictive SENsorimotor Control and Emulation'.

PRESENCE is founded on a model of interaction in which an actor has in mind the *needs* [50] of an observer and an observer has in mind the *intentions* of an actor, and that both achieve these behaviors by emulating each other. This is a crucial difference between PRESENCE and contemporary architectures for speech-based human-machine interaction. PRESENCE thus provides a fundamental mechanism for supporting *communicative* behavior between participants in an ongoing dialogue, and breaks away from the traditional stimulus-response model of the speech chain towards a more integrated view based on phase-locked control loops.

The notion of hierarchical feedback control, as posited by PCT, is assumed to be inherent in such interactive behavior, and this is supported by the observation that it provides a credible computational mechanism to underpin Lindblom's 'H&H' theory of speech [37]. H&H describes the process whereby speakers exhibit real-time control of their articulatory 'effort' in order to balance the needs of communicative effectiveness against the energy demands involved in speaking. Lindblom has hypothesized that such a mechanism could not only explain the apparent random variability that is observed in speech, but also that it provides a framework that would support the emergence of a *contrastive* system of communicative behavior (i.e. the evolution of the phonemic structure of spoken language). These behaviors are fundamental to the special nature of speech, language, and communication, and yet they are completely missing from contemporary systems for speech-based human-machine interaction.

PRESENCE not only incorporates the principles of H&H, but also extends them to cover both the production and perception of spoken language within a communicative context, based on a

general recursive framework for simulating and predicting both speaker and listener behavior. One of its fundamental tenets is that performance benefits should accrue from maintaining a close connection between the hitherto independent processes of speech input and speech output. At the practical level, this could simply mean the sharing of models between recognition and synthesis. However, the implications run much deeper - PRESENCE implies that the process of spoken language generation/synthesis should be invoked *as part of* the process of spoken language recognition/understanding, and that the process of spoken language recognition/understanding should be invoked *as part of* the process of spoken language generation/synthesis.

A. The communicative loop

Somewhat surprisingly, PRESENCE dictates that the primary function of a speech-based system is *not* to speak, or to listen, but to interact with a user in order to meet the *system's* needs. This latter point might appear counterintuitive in that it would seem that the needs of a user should be paramount. However, in order for a system to serve the interests of the user, the needs of the system have to be declared in terms of meeting those user's needs. Indeed, PRESENCE predicts that it is only by establishing such basic drives *explicitly* that it is possible to design an automated system that would perform any behavior at all.

A basic communicative loop is illustrated in Fig. 3. The needs of the system are given as a prior (by the system designer) and are specified as a multidimensional reference vector $S{:}n_i$ where S indicates 'system', n signifies 'needs', and i is an index over I independent system needs. For example, a system may be configured to complete a transaction to a specified level of quality within a certain period of time. The needs of the user are to be determined by the system and are specified as a multidimensional reference vector $U{:}n_j$ where U indicates 'user', n

signifies ‘needs’, and j is an index over J independent user needs. If the system’s needs and the user’s needs are aligned, then the resulting communicative behaviors are likely to be both effective and efficient. However, if the two sets of needs are in some sense conflicting (for example, a user may wish to maintain the engagement for as long as possible), then the resulting interaction might exhibit classic symptoms of an unstable control system such as oscillatory behavior or even hard limiting. Clearly both system and user drives are a function of the communicative context – the application - and hence system drives require careful thought on the part of the system designer: what is appropriate for an automated enquiry service may be quite inappropriate for a robot companion.

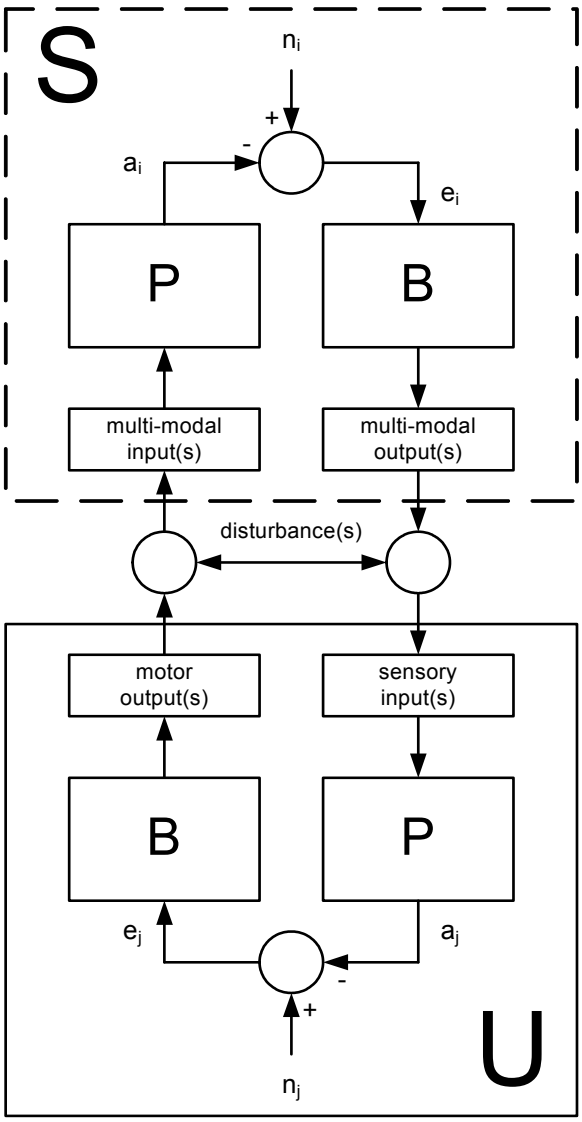

Fig. 3. Illustration of the basic communicative loop in the PRESENCE architecture.

The basic operation of the communicative loop illustrated in Fig. 3 is as follows: At the highest level, a perceptual process in the system *S:P* determines, for each of its key criteria, the current state of achievement $S{:}a_i$, and the difference between $S{:}a_i$ and $S{:}n_i$ produces an error signal $S{:}e_i$ that drives system behavior *S:B* in such a way as to reduce $S{:}e_i$. In parallel, a perceptual process in the user *U:P* determines, for each of their key criteria, the current state of achievement $U{:}a_j$, and the difference between $U{:}a_j$ and $U{:}n_j$ produces an error signal $U{:}e_j$ that drives the user's behavior *U:B* in such a way as to reduce $U{:}e_j$. The result is that the system will

consistently act to overcome any obstacle or disturbance (whether it is user–generated, system-generated and/or present in the operating environment) that interferes with the realization of its intended behavior, whilst at the same time the user is acting in order to realize their intended behavior.

In order for either the system or the user to maximize their achievements in the context of the interaction, it is clearly necessary for each to determine the needs of the other. In particular, a cooperative system would require its needs $S{:}n_i$ to be expressed as a function of its understanding of its user's needs $U{:}n_j$, and a cooperative user would express their needs $U{:}n_j$ in terms of the needs of the system $S{:}n_i$. Such a recursive arrangement not only facilitates an alignment between the behavior of the two participants, but it also allows both to achieve success in otherwise *unpredictable* circumstances. Again, any mismatch between system and user, e.g. arising from one participant misunderstanding the needs of the other, is destined to lead to communicative difficulties and/or failure.

Determining a user's needs $U{:}n_j$ is achieved in PRESENCE either by access to a predictive model/emulation of the user $S{:}E(U{:}n_j)$ (in which case a user's needs may be given as a prior, or they may need to be estimated by running a simulation of the user), by recognizing the user's expressed needs $S{:}P(U{:}B(U{:}n_j))$, or by requesting the user to express their needs $S{:}B(U{:}B(U{:}n_j))$. The choice of which of these strategies to pursue at any given point in time would depend on the output of an emulation process aimed at predicting and assessing the possible outcomes against constraints conditioned on other high level system needs (such as meeting a user's needs within a certain time frame). So, for example, a successful system might be one which could accurately anticipate the needs of a user based on minimal interaction in order to satisfy them in the shortest possible time.

Since the achievements of the system $S{:}a_i$ are expressed in terms of meeting user needs $U{:}n_j$, then $S{:}a_i$ is actually determined by $U{:}e_j$ – the user's internal error signal. This means that a secondary function of the PRESENCE architecture is an ability, not only to determine a user's needs, but also to estimate the degree to which those needs are being met (by the system). In other words, the self-evaluation question "how well am I doing?" (from the system point of view) is *implicit* in the PRESENCE architecture. Interestingly, the degree to which a user's needs are being met $U{:}e_j$ will have a direct influence on user behavior $U{:}B$ and may actually be manifest in the form of emotion. Hence, PRESENCE predicts that the effectiveness of human-machine interaction would be greatly enhanced if the system was able to recognize such user behavior. PRESENCE also predicts that interaction would also be enhanced if a system was able to communicate the degree to which its needs are being met $S{:}e_i$ through appropriate system behavior $S{:}B$.

Therefore, perhaps surprisingly, it can be seen that the main building block in the PRESENCE architecture is not a set of low-level sensorimotor processes concerned with detailed acoustic-phonetic speech recognition and synthesis behavior. Rather, it is a high-level communicative loop structure that is conditioned on the fundamental purpose of the overall system; its essential drives being derived from what would be regarded as the application 'back-end' in a conventional architecture. It is this high-level communicative loop that drives and shapes the low-level communicative behaviors – not the other way around. Of course spoken language recognition and generation are essential features of this high-level structure, but what is being recognized and generated is not yet specified in acoustic or even linguistic terms; at this level the constructs are communicative locutionary acts that form key steps within an iterative optimization process (targeted at the meeting of system and user needs in an efficient and timely

manner).

B. Internal structure

The internal structure of the PRESENCE architecture is illustrated in Fig. 4 (adapted from [6]). This is a general-purpose view intended to represent both a model of human behavior as well as a putative functional architecture for a practical system. This duality of purpose explains why there are not only references to universal concepts such as action and interpretation, but there are also references to motivational and emotional parameters as well as notions such as attention. Again, the suggestion that such human-like behaviors might have a key role to play in the design of an artificial cognitive system is made immediately apparent by the PRESENCE architecture.

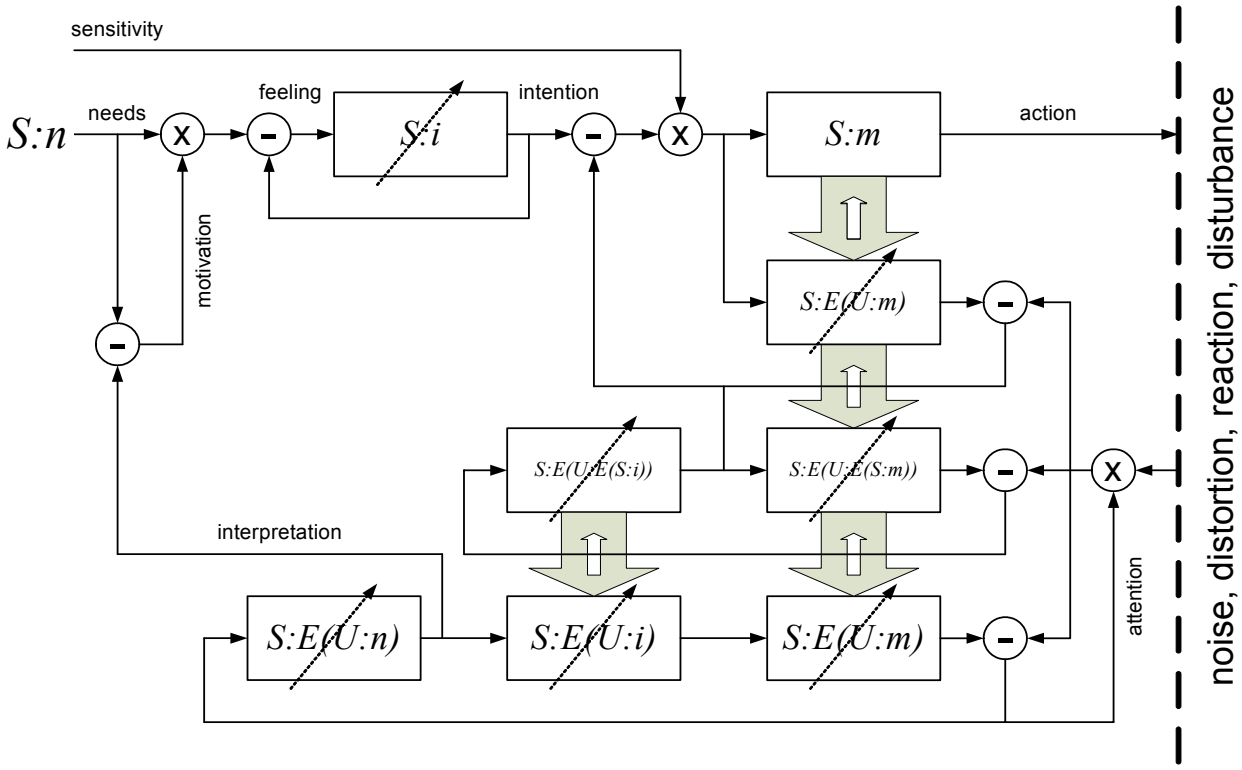

Fig. 4. Internal structure of the PRESENCE architecture.

The architecture shown in Fig. 4 is organized into four layers. The top layer is the main path for motor behavior such as speaking. A system's needs *S:n* modulated by motivation, causes the

selection of a communicative intention *S:i* that would satisfy those needs. The selection mechanism can be implemented as a search process, and this is indicated by the diagonal arrow running through the *S:i* module. The selected intention drives both actual motor behavior *S:m* and an emulation of possible motor behavior *S:E(S:m)* on the second layer. Sensory input feeds back into this second layer, providing a check as to whether the desired intention has been met. If there is a mismatch between intended behavior and the perceived outcome, then the resulting error signal will cause the system to alter its behavior appropriately.

The third layer of the model captures the empathetic relationship between system as a speaker and the user as a listener that conditions the speaking behavior of the system. *U:E(S:i)* represents the emulation by the user of the intentions of the system, and *S:E(U:E(S:i))* represents the emulation of that function by the system. A similar arrangement applies to *S:E(U:E(S:m))* – the system's emulation of the user's emulation of the systems motor output. The fourth layer represents the system's means for interpreting the needs, intentions and behavior of a user though a process of emulating the user's needs *S:E(U:n)*, intentions *S:E(U:i)* and behavior *S:E(U:m)*.

The second, third and fourth layers are able to exploit the information embedded in the previous layers, and this is indicated in Fig. 4 by the large block arrows. This process is equivalent to parameter sharing between the different models and thus represents not only an efficient use of information but also offers a mechanism for learning. In fact such a process may be bi-directional, and the potential flow of information in the opposite direction is indicated in Fig. 4 by the small block arrows.

C. Recursive nesting

Clearly, the basic communicative loop in the PRESENCE architecture contains system

components that are themselves realized using similarly-structured building blocks. The PRESENCE architecture is thus inherently recursive and therefore hierarchical in structure, with further refinements in behavior coming from the operation of the nested components.

For example, if a system is driven to ask a question, then a nested structure is required in order to determine what is actually going to be said. The communicative loop at this level would use emulation mechanisms to take in to consideration the possible consequences of particular choices of linguistic content, and such emulators would effectively be simulations of the user's mirror understanding system - 'synthesis-by-analysis'. The significance of such a structure is that it provides a natural mechanism for allowing the resulting linguistic constructs to be truly communicative, i.e. words and phrases would be chosen specifically to maximize the effectiveness of the communication, avoiding confusion and enhancing clarity in the context of the ongoing interaction – all these features being estimated using forward models to perform predictive emulation of the possible *consequences* of the proposed linguistic output.

Similarly, interpretation of a user's response would be based on reference to the system's own generative capabilities – 'analysis-by-synthesis'. Again, the significance of such a structure is that it provides a natural mechanism for accessing the 'hidden' meaning of user behavior by virtue of placing the system in the virtual position of the user; understanding arises as an emergent by-product of the synchronization of the knowledge and beliefs of both system and user – the system *empathizes* with the user in the widest sense.

D. Speech-based interaction

As outlined thus far, the PRESENCE architecture is somewhat neutral with respect to the modalities of a system's interaction with a user. In fact this is a real bonus, since it means that multi-modal behavior is treated as the general case of communicative human-machine

interaction. PRESENCE clearly provides a mechanism whereby a system may make choices between different communicative modalities based on an understanding of the differing characteristics of the individual channels (e.g. in terms of information transfer rates, memorability, noise, interference etc.) and the projected implications of their use at each point in an interaction. That is, a system may itself choose to use a particular modality based on its estimation of the effectiveness of that strategy in meeting its needs – and this may change in the course of an ongoing interaction.

However, the real power of the PRESENCE architecture becomes clear when considering speech-based human-machine interaction. Of course, language in itself provides a higher bandwidth channel than any other modality (~50 bits-per-second [51]), but spoken language approximately doubles that through the addition of expressive behavior that carries further linguistic information such as prosody, as well as para-linguistic information such as individuality and expression. Such behaviors present major problems for state-of-the-art speech-based human-machine interaction, but they are seen as *central* to the functionality of PRESENCE.

In practice, this means that not only would a system based on the PRESENCE architecture be able to choose its words carefully, but it would also be capable of adjusting its pronunciation in order to avoid potential confusion and to overlay expressive behavior appropriate to its internal states or the needs of the communication (e.g. it would automatically start to speak louder in a noisy environment). These behaviors emerge because of the system's ability to emulate the user and hence accommodate the user's expectations based on an estimation of their listening experience. In other words, in PRESENCE the process of speech generation is mediated by reference to a feedback path involving speech recognition – the system would effectively be

listening to its own output (either overtly or using an 'inner loop' [52]) in order to judge whether it was having the intended effect on the listener, a concept referred to by the author as 'reactive speech synthesis'.

Likewise, the system would be able to interpret the intention behind a user's particular choice of words and pronunciation, and the implications of expressive behavior, all by the means of reference to low-level emulations of mirrored sensorimotor structures. In other words, in PRESENCE the process of speech recognition is mediated by reference to a forward model based on the emulation of speech generation – the system would effectively be determining the implications of what is being said by *imagining* itself saying it.

These features of the PRESENCE architecture mean that such a system would exhibit communicative behavior in both the production and perception of speech. Issues such as pronunciation modeling are sidestepped because the predictive feedback control structures compensate automatically for the communicative context. PRESENCE thus provides an opportunity for a real advance towards 'intelligent' behavior in speech-based human-machine interaction.

IV. Towards Practical Systems

The foregoing sections have argued the case for viewing the production and perception of spoken language within a single theoretical framework - PRESENCE. From this new perspective it is immediately possible to extract some practical implications for future spoken language technology.

For example, PRESENCE suggests an architecture for a new type of *reactive* speech synthesizer that would actively modify its output behavior as a function of its perceived effectiveness – talking louder in a noisy environment and actively altering its pronunciation to

maximize intelligibility and minimize potential confusion. In order to achieve such advanced behavior, PRESENCE indicates that such a synthesizer would need to include a model of the listener within the feedback loop, and this would be achieved by simulating the behavior of the listener using an automatic speech recognizer. This means that such a system could effectively be described as 'synthesis-by-recognition' (SbR). As far as the author is aware, no contemporary text-to-speech synthesizer utilizes this kind of feedback, although something along these lines was suggested by Fallside some time ago [53] and a related scheme is currently being used to train a low-level speech synthesizer to imitate speech [54].

For recognition, PRESENCE suggests an architecture that incorporates an emulation of the speaker, i.e. a generative model of speech whose output is compared with incoming speech data. Of course, almost all state-of-the-art ASR systems already employ generative models in the form of hidden Markov models, so the conventional approach to ASR would already appear to fit nicely within the PRESENCE framework. In some sense this is correct, however as outlined earlier, a standard HMM is typically a rather poor model of speaker behavior. To fully realize the opportunities offered by PRESENCE, it is necessary to invoke a new type of architecture for speech recognition that, instead of HMMs, would incorporate a generative model that is closer to an actual speech synthesizer in order to perform 'recognition-by-synthesis' (RbS). In fact such an idea was proposed over 20 years ago by Bridle and Ralls [55] and since that time a number of researchers, inspired in part by the Motor Theory of speech perception [56], have been attracted to the prospect of incorporating models of speech production within automatic speech recognition [57]-[59]. However, the difference between such approaches and the one being proposed here is that the neuroscience studies underpinning this aspect of the PRESENCE architecture suggest that such an embedded model of speech generation should be derived, not

from the voice of the speaker, but from the voice of the listener (which, in this case, is a machine!).

This apparent dilemma points the way towards the need to unify research into automatic speech recognition with research into text-to-speech synthesis. Not only does PRESENCE suggest an architecture within which each refers to the other, but this leads to a powerful recursive structure in which it is possible to envisage 'recognition-by-synthesis-by-recognition' (RbSbR), 'synthesis-by-recognition-by-synthesis' (SbRbS) and so on, with each layer providing greater fidelity and refinement than the layer above.

V. Experimental Work

Clearly the implications of PRESENCE for the architecture of future spoken language systems are far reaching in both scope and potential impact. By integrating both speech recognition and generation within a single recursive structure for speech-based interaction, PRESENCE posits a very different approach to system design and implementation. This means that it is quite difficult to exploit the traditional experimental framework for developing a spoken language system, since the conventional approach involves the bottom-up instantiation of independent system components - the very components that PRESENCE seeks to integrate. Therefore, in marked contrast to such familiar methodologies for system construction, PRESENCE points towards a more top-down design methodology, starting with the definition of a system's basic needs embedded within a high-level interactive control structure.

Therefore, in order to lend some experimental support to the novel architecture proposed in this paper, a preliminary investigation has been conducted into a physical instantiation of high-level acoustic interaction between a robot and a human being. This approach was chosen as the fastest means for demonstrating the essential principles of the overall PRESENCE architecture

without using simulation or approximation.

The task selected was to create an embodied device that could learn to produce motor behavior in time to rhythmic input (much like someone clapping along to music). This might appear to be a long way from something like automatic speech recognition, however the sophisticated coordination and synchronization of behavior between system and user in speech-based human-machine interaction is just the kind of problem that remains a challenge for contemporary approaches. PRESENCE, on the other hand, offers an immediate solution based on interlocking control structures.

A. The robot

A humanoid robot - ALPHA REX - was constructed using the LEGO® MINDSTORMS® NXT platform [60]. The device consisted of a central 32-bit microprocessor controller, three interactive servo motors and four sensors; sound, light, touch and ultrasonic. The robot was programmed by USB connection to a PC using the standard LEGO® MINDSTORMS® NXT software environment. Since the aim was to demonstrate the core principles of PRESENCE, the 'drive' of the robot was declared as a high-level 'need' to maximize synchrony between its own behavior and that of an external source. The resulting software architecture instantiated this need as three parallel sub-loops within the overall control loop; first, a loop to generate its own rhythmic behavior; second, a loop to sense its own behavior; and third, a loop to sense any external behavior. In this first implementation, the second loop was embedded in the first. However, in order to control a genuine clapping response in a future implementation, this loop would need to be instantiated independently.

The sensor and motor sub-loops each generated a variable that represented the precise time of each 'tap' (i.e. the robot's tap and the external tap), and the overall control loop compared the

two times and generated an error signal that was proportional to the difference. The error signal then increased or decreased the robot's rhythmic behavior until it matched the external source. The outcome is thus that the robot learns to adapt its behavior until there is synchrony between its behavior and that of the external source.

B. Results

Although this work represents a relatively simple example of PRESENCE-based architecture, it nevertheless successfully demonstrated an alternative to the traditional stimulus-response (S-R) view of intelligent behavior. Unlike an S-R model, the robot did not suffer from internal system delays that would give rise to behavior with the same rhythm but out of synchronization. Nor did the robot need to compute complex analytical solutions in order to estimate such delays. In other words, the implementation of a PRESENCE-like structure gave the robot an ability to 'anticipate' the external behavior, and this was evidenced by the fact that the robot always made one more action precisely at the appropriate moment even though the external behavior had ceased.

Fig. 5 illustrates this behavior for a spoken input. The experimenter (U) uttered a short sequence consisting of the words "one" and "two" spoken at regular intervals, and the robot (S) generated 'taps'. As can be seen from the Figure, the robot starts to tap by the third user utterance and gradually established a rhythm that is precisely synchronized with the onset of the eighth and ninth utterances. The ninth utterance is the last, but the robot emits one final tap at the time at which the next word would have been expected.

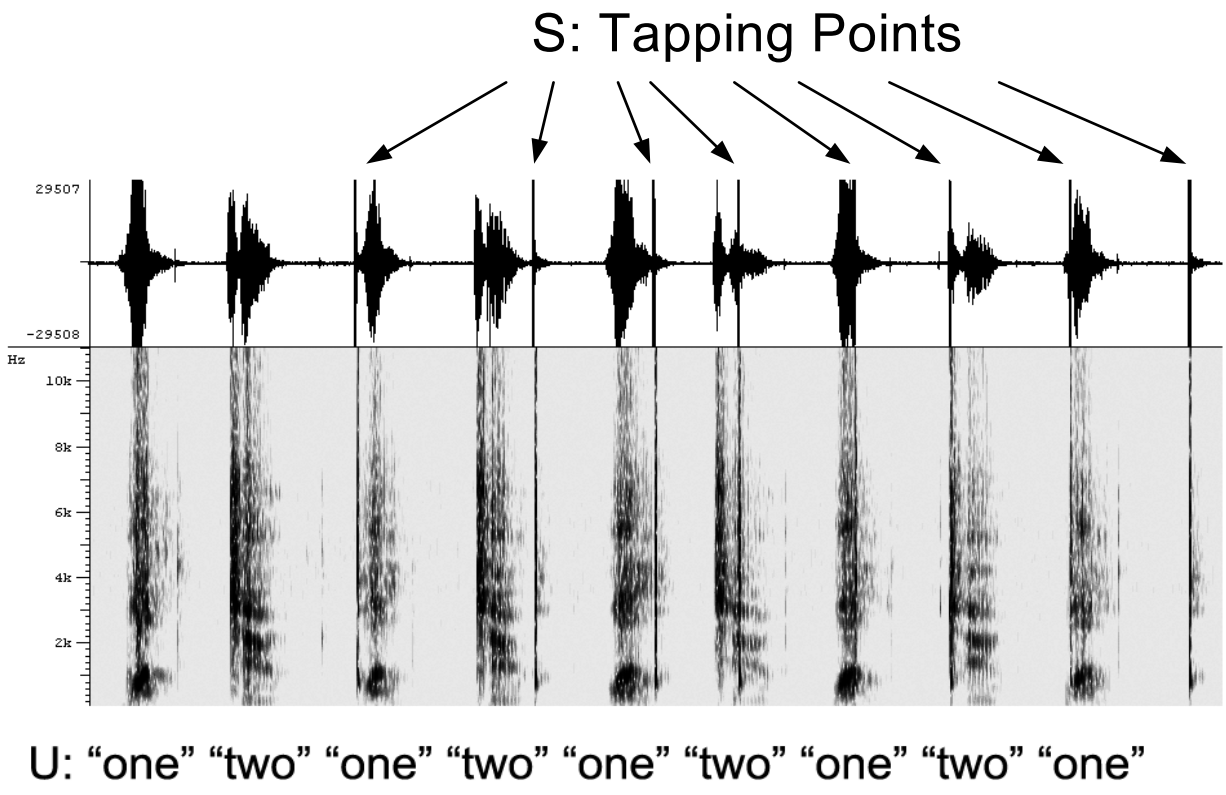

Fig. 5. Illustration of the synchronization of robot behavior with spoken input.

VI. Discussion

A. System initialization

A key difference between PRESENCE and a more conventional architecture for speech-based human-machine interaction is that PRESENCE can be said to ‘know’ what it is saying and why it is saying it. As a consequence, it also ‘knows’ what a user is saying and why they are saying it. PRESENCE achieves this feat by providing a framework - inspired by insights into the neurobiology of living systems - that unifies the processes of generating and interpreting behavior. However, a major issue is just how such a framework is established in the first place. It is one thing to hypothesize a general recursive structure of the type described, but at some point the parameters of a particular system need to be specified. In other words, if the generation of

appropriate behavior is controlled by interpretation of that behavior, and the interpretation of behavior is made with respect to a putative generator, then there appears to be a fundamental lack of 'grounding' within the system.

This apparent dilemma goes away if one is concerned with living systems, since the grounding is provided by the physical attributes of the individual organisms and the implicit commonality of those physical attributes between different members of the same species. However, an automated system shares very little physical structure with a human user.

This suggests three possibilities: first, priority could be given to research into speech generation techniques that mimic more closely the physical human speech production process [61] (such research is no longer 'in vogue', yet it may be crucial to the development of the next generation of speech-based interactive systems); second, it may be necessary to create systems that are able to acquire the necessary grounding by learning the appropriate skills in a situated and embedded environment (i.e. analogous to the process by which infants acquire social and linguistic skills) [62]; third, perhaps it will *never* be possible to establish truly effective speech-based human-machine interaction (in much the same way that speech-based human-*animal* communication is fundamentally limited by a lack of a shared frame of reference).

B. Recognizing users

In the system description presented above, it was taken for granted that users would present themselves to the system in a cooperative and friendly manner. However, in a real-world application environment - especially for tasks those that do not involve a 'captured' user (such as a telephone-based system) - even this simple assumption may be invalid. Therefore, in the general case, a system might have to take control of a range of different scenarios. For example, it might be necessary to be able to recognize the presence of a user, to identify a user in a

complex environment, or to discriminate between users and non-users. The 'needs' structure inherent in the PRESENCE architecture provides a mechanism for handling such cases, and it could even be invoked to actively search for users – a behavior that currently has to be explicitly 'programmed in' to a speech-enabled robot [63].

C. The particulate structure of language

One of the disadvantages of the conventional approach to speech-based human-machine interaction is that each additional component effectively adds a new set of parameters that have to be estimated from training data (or learnt during the operation of the system). As a consequence, the number of conditional dependencies within the overall system grows exponentially with the complexity of the system, as does the required amount of training data. In marked contrast, by virtue of its inherent recursive structure, free variables are factored in the PRESENCE architecture. Not only does this provide an efficient mechanism for storing information and for maximizing the value of limited training material, but it reflects the particulate structure of a self-diversifying system such as language [64]. It seems that speech and language have evolved precisely to exploit such efficiencies, and hence have given rise to a communicative medium with vast expressive potential based on a physical system – the human vocal apparatus – possessing relatively few degrees of freedom.

D. Knowledge re-use

It may appear that the creation of PRESENCE is aimed at discarding much of the good research into speech-based human-machine interaction that has already taken place. However, by virtue of the fact that the speech technology community has been facing such a difficult challenge over many years, scientists and engineers already have at their disposal a wide range of very powerful tools for computational modeling. Many of the processes embedded within the

PRESENCE architecture – pattern matching, sequential search, predictive models – are already well understood; what is new is the conceptual framework within which such processes are embedded. In the same way that algorithms from speech processing have pervaded other areas of pattern processing, so too the advanced computational processes required within PRESENCE might well serve to influence the wider scientific community studying the neurobiology of living systems.

VII. Conclusion

In response to the hypothesis that the quantity of training data required to improve state-of-the-art speech-based human-machine interaction seems to be growing exponentially, and that performance is asymptoting to a level that may be inadequate for many real-world applications, this paper has presented the outline of a novel architecture that has been inspired by recent findings in the neurobiology of living systems. Called PRESENCE - 'PREdictive SENsorimotor Control and Emulation' – this new architecture blurs the distinction between the components of a traditional spoken language dialogue system and, instead, focuses on a recursive hierarchical feedback control structure driven by high-level system needs. Cooperative and communicative behavior emerges as a by-product of an architecture that is founded on a model of interaction in which the system has in mind the needs and intentions of a user, and a user has in mind the needs and intentions of the system.

Much detail has yet to be worked out, yet it is clear that the implications of this new architecture are potentially far reaching. Not only might PRESENCE provide a means to construct more effective speech-based human-machine interfaces based on an emergent natural intelligence [65] but, for example by drawing on recent work by Oztop et al [66], it also offers the possibility of creating biologically credible computational models of human spoken language

behavior, thereby unifying the currently divergent fields of speech science and technology [67]. Indeed, such a convergence of knowledge and disciplines is already being fostered by the newly emerging transdisciplinary field of 'Cognitive Informatics' [68]. Cognitive Informatics aims to forge links between a diverse range of disciplines spanning the natural and life sciences, informatics and computer science, and is founded on the conviction that many fundamental questions of human knowledge (such as spoken language processing) share a common basis - an understanding of the mechanisms of natural intelligence and the cognitive processes of the brain. The appearance of Cognitive Informatics and its community of like-minded researchers presents a unique opportunity for research into speech-based human-machine interaction to sit at the very heart of this new field [69].

ACKNOWLEDGMENT

The author would like to thank Dr. Peter Wallis for his suggestion to use a clapping robot as a challenging interaction scenario, and the reviewers for their helpful and insightful comments.

REFERENCES

[1] R. K. Moore, "Research challenges in the automation of spoken language interaction," Keynote talk, *Proc. COST278 and ISCA Tutorial and Research Workshop (ITRW) on Applied Spoken Language Interaction in Distributed Environments* (ASIDE 2005), Aalborg University, Denmark, 2005.

[2] R. K. Moore, "Modelling data entry rates for ASR and alternative input methods", *Proc. INTERSPEECH 2004 ICSLP*, Jeju, Korea, 2004.

[3] F. Jelinek, "Five speculations (and a divertimento) on the themes of H. Bourlard, H. Hermansky, and N. Morgan," *J. Speech Communication* 18, 1996, pp. 242-246.

[4] R. K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners," *Proc. EUROSPEECH'03*, Geneva, 2003 , pp. 2582-2584.

[5] E. Keller, "Towards Greater Naturalness: Future Directions of Research in Speech Synthesis," in *Improvements in Speech Synthesis*, Keller, E., Bailly, G, Monaghan, A., Terken, J. and Huckvale, M. (eds.), Wiley & Sons, Chichester, UK, 2001.

[6] R. K. Moore, "Spoken language processing: piecing together the puzzle," *J. Speech Communication*, 49, 2007, pp. 418-435, http://dx.doi.org/10.1016/j.specom.2007.01.011.

[7] A. Newell, J. Barnett, J. Forgie, C. Green, D. Klatt, J. Licklider, J. Munson, R. Reddy and W. Woods, *Speech Understanding Systems*, North-Holland/American Elsevier, 1973.

[8] D. Klatt, "Review of the ARPA speech understanding project," *J. Acoustical Soc. of America*, 62, 1977, pp. 1345-1366.

[9] B. T. Lowerre, *The HARPY Speech Recognition System*, PhD Thesis, Dept. Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1976.

[10] H. Hermansky, 1998. "Should recognizers have ears?" *Speech Communication* 25, 1998, pp. 3-27.

[11] C-H. Lee, "From decoding-driven to detection-based paradigms for automatic speech recognition," *Proc. Int. Conf. on Spoken Language Processing* (ICSLP), Jeju, Korea, 2004.

[12] O. Scharenborg O, V. Wan and R. K. Moore, "Capturing fine-phonetic detail in speech through automatic classification of articulatory features," *Proc. ISCA workshop on Speech Recognition and Intrinsic Variation*, Toulouse, 20 May 2006.

[13] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard, and M. Athineos, "Pushing the envelope - aside," *IEEE Signal Processing Magazine* 22 (5), 2005, pp. 81-88.

[14] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to segment models: a unified view of stochastic modeling for speech recognition.," *IEEE Trans. Acoustics, Speech and Signal Processing* 4, 1996, pp. 360-378.
[15] J. Sun and L. Deng, "An overlapping-feature based phonological model incorporating linguistic constraints: Applications to speech recognition," *J. Acoustical Soc. of America* 111(2), 2002, pp. 1086-1101.
[16] O. Scharenborg, L. ten Bosch, L. Boves, and D. Norris, "Bridging automatic speech recognition and psycholinguistics: extending Shortlist to an end-to-end model of human speech recognition," *J. Acoustical Soc. of America* 114(6), 2003, pp. 3023-3035.
[17] O. Scharenborg, D. Norris, L. ten Bosch, and J. M. McQueen, "How should a speech recogniser work?" *Cognitive Science* 29, 2005, pp. 867-918.
[18] V. Maier and R. K. Moore, "An investigation into a simulation of episodic memory for automatic speech recognition," *Proc. INTERSPEECH 2005*, Lisbon, 2005.
[19] K. Kirchhoff and S. Schimmel, "Statistical properties of infant-directed vs. adult-directed speech: insights from speech recognition," *J. Acoustical Soc. of America* 117(4), 2005, pp. 2224-2237.
[20] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication* 22, 1997, pp. 1-16.
[21] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, "Premotor cortex and the recognition of motor actions," *Cognitive Brain Research* 3, 1996, pp. 131-141.
[22] G. Rizzolatti and L. Craighero, "The mirror-neuron system," *Annual Review of Neuroscience* 27, 2004, pp. 169-192.
[23] M. Wilson and G. Knoblich, "The case for motor involvement in perceiving conspecifics," *Psychological Bulletin* 131 (3), 2005, pp. 460-473.
[24] V. Gallese, C. Keysers and G. Rizzolatti, "A unifying view of the basis of social cognition," *Trends in Cognitive Science*, vol.8, no.9, 2004, pp. 396-403.
[25] C. Frith. "Attention to action and awareness of other minds," *Consciousness and Cognition* 11, 2002, pp. 481-487.
[26] G. Rizzolatti and M. A. Arbib, "Language within our grasp," *Trends in Neuroscience* 21, 1998, pp. 188-194.
[27] F. Aboitiz, R. Garcia, E. Brunetti and C. Bosman, "Imitation and memory in language origins," *Neural Networks* 18, 2005, pp. 1357.
[28] E. Kohler, C. Keysers, M. Allessandra Umilta, L. Fogassi, V. Gallese and G. Rizzolatti, "Hearing sounds, understanding actions: action representation in mirror neurons," *Science* 297, 2002, pp. 846-848.
[29] F. Pulvermüller, "Brain mechanisms linking language and action," *Nature Neuroscience Review* 6, 2005, pp. 576-582.
[30] L. Fadiga, L. Craighero, G. Buccino and G. Rizzolatti, "Speech listening specifically modulates the excitability of tongue muscles: a TMS study," *European J. of Neuroscience* 15, 2002, pp. 399-402.
[31] S. M. Wilson A. P. Saygin, M. I. Sereno and.M. Iacoboni, "Listening to speech activates motor areas involved in speech production," *Nature Neuroscience* 7(7), 2004, pp. 701-702.
[32] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition* 21, 1985, pp. 1-36.
[33] P. B. Denes and E. N. Pinson, *The Speech Chain: The Physics and Biology of Spoken Language*, New York: Anchor Press, 1973.
[34] W. T. Powers, *Behaviour: The Control of Perception*, Hawthorne, NY: Aldine, 1973.
[35] M. M. Taylor, P. S. E. Farrell and J. G. Hollands, "Perceptual control and layered protocols in interface design: II. The general protocol grammar," *Int. J. Human-Computer Studies* 50, 1999, pp. 521-555.
[36] J-C. Junqua, "The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication* 20, 1996, pp. 13-22.
[37] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," in: W. Hardcastle and A. Marchal (Eds.), *Speech Production and Speech Modeling*, Kluwer, 1990, pp. 403-439.
[38] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature Reviews: Neuroscience* 5, 2004, pp. 831-843.
[39] R. Grush, "The emulation theory of representation: motor control, imagery, and perception," *Behavioral and Brain Sciences* 27, 2004, pp. 377-442.
[40] R. Grush, "Perception, imagery, and the sensorimotor loop," in: *A Conciousness Reader*, Esken and Heckmann (eds), Schoeningh Verlag, 1998.
[41] S. J. Cowley, "Simulating others: the basis of human cognition," *Language Sciences* 26, 2004, pp. 273-299.
[42] V. G. J. Gerdes and R. Happee, "The use of an internal representation in fast goal-directed movements: a modeling approach," *Biological Cybernetics* 70, 1994, pp. 513-524.
[43] M. Studdart-Kennedy, "Mirror neurons, vocal imitation, and the evolution of particulate speech," in: M.I. Stamenov, V. Gallese (Eds.), *Mirror Neurons and the Evolution of Brain and Language*, Philadelphia: Benjamins, 2002, pp. 207–227.
[44] M. Meltzoff and K. Moore, "Explaining facial imitation: a theoretical model," *Early Development and Parenting* 6, 1997, pp. 179-192.
[45] J. Hawkins, *On Intelligence*, Times Books, 2004.
[46] V. B. Mountcastle, "An organizing principle for cerebral function: the unit model and the distributed system," In: *The Mindful Brain*, G.M. Edelman, V.B. Mountcastle (Eds.), MIT Press. 1978.
[47] J. Hawkins and D. George, "Hierarchical temporal memory," Technical White Paper: Numenta Inc., 2006.
[48] F. Wörgötter and B. Porr, "Temporal sequence learning, prediction, and control: a review of different models and their relation to biological mechanisms," *Neural Computation* 17, 2005, pp. 245-319.
[49] A. G. Barto, "Adaptive critics and the basal ganglia," in: *Models of Information in the Basal Ganglia*, J.C. Houk, J. Davis and D. Beiser (Eds.), MIT Press, Cambridge MA, 1995, pp. 215-232.
[50] A. H. Maslow, "A theory of human motivation," *Psychological Review* 50, 1943, pp. 370-396.
[51] C. Cherry, *On Human Communication: A Review, a Survey and a Criticism*, The MIT Press, 1978.
[52] W. J. M. Levelt, "Monitoring and self-repair in speech," *Cognition* 14, 1983, pp. 41-104.
[53] F. Fallside, "Synfrec: speech synthesis from recognition using neural networks," *Proc. ESCA Workshop on Speech Synthesis*, 1990, pp. 237-240.
[54] I.S. Howard and M. A. Huckvale, "Training a vocal tract synthesizer to imitate speech using distal supervised learning," *Proc. SPECOM*, 2005, pp. 159-162.
[55] J. S. Bridle and M. P. Ralls, "An approach to speech recognition using synthesis by rule," in: F. Fallside and W. Woods (Eds.), *Computer Speech Processing*, Prentice Hall, 1985.
[56] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition* 21, 1985, pp. 1-36.

[57] M. Blomberg, R. Carlson, K. Elenius, B. Granström, S. Hunnicutt, R. Lindell and L. Neovius, "Speech recognition based on a text-to-speech synthesis system," in J. Laver and M. A. Jack (Eds.), *European Conference on Speech Technology*, Edinburgh, 1987, pp. 369-372.
[58] R. K. Moore, "Critique: the potential role of speech production models in automatic speech recognition," *Journal of the Acoustical Society of America* 99(3), 1996, pp. 1710-1713.
[59] E. McDermott and A. Nakamura, "Production-oriented models for speech recognition," *IEICE Trans. Inf. & Sys*. E89-D(3), 2006, pp. 1006-1014.
[60] LEGO MINDSTORMS, http://www.mindstorms.com
[61] R. Hofe, research homepage, http://www.dcs.shef.ac.uk/~robin/
[62] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science* 26, 2002, pp. 113-146.
[63] A. Drygajlo, P. Prodanov, G. Ramel, M. Meisser and R. Siegwart, "On developing voice enabled interface for interactive tour-guide robots," *J. Advanced Robotics*, Robotics Society of Japan, 2003.
[64] W. L. Abler, "On the particulate principle of self-diversifying systems," *J. Social Biol. Struct.* 12, 1989, pp. 1-13.
[65] A. Turing, "Computing machinery and intelligence," *Mind*, LIX(236), 1950, pp. 433-460.
[66] E. Oztop, M. Kawato and M. Arbib, "Mirror neurons and imitation: a computationally guided review", *Neural Networks* 19, 2006, pp. 254-271.
[67] R. K. Moore, "Towards a unified theory of spoken language processing," *Proc. 4th IEEE International Conference on Cognitive Informatics*, Irvine, CA, USA, 2005.
[68] Y. Wang, "On cognitive informatics," *Brain and Mind* 4, 2003, pp. 151-167.
[69] R. K. Moore, "Cognitive informatics: the future of spoken language processing?," Keynote talk, Proc. SPECOM - *10th Int. Conf. on Speech and Computer*, Patras, Greece, 2005.

Roger K. Moore (M'00) was born in Swanage, Dorset, UK in 1952. He studied Computer and Communications Engineering at the University of Essex and was awarded the B.A. (Hons.) degree in 1973. He subsequently received the M.Sc. and Ph.D. degrees from the same university in 1975 and 1977 respectively. After a period of post-doctoral research in the Phonetics Department at University College London, in 1980 Prof. Moore established the speech recognition research team at the Royal Signals and Radar Establishment (RSRE) in Malvern. In 1985 Prof. Moore became head of the newly created 'Speech Research Unit' (SRU) and subsequently rose to the position of Deputy Chief Scientific Officer in the 'Defence and Evaluation Research Agency' (DERA). Following the privatization of the SRU in 1999, Prof. Moore assumed the role of Chief Scientific Officer at 20/20 Speech Ltd. until 2004 when he was appointed Professor of Spoken Language Processing at the University of Sheffield. Prof. Moore has authored and co-authored over 100 scientific publications in the general area of speech technology applications, algorithms and assessment. He is a Fellow of the UK Institute of Acoustics and a Visiting Professor in the Department of Phonetics and Linguistics at University College London. He is Editor-in-Chief for 'Computer Speech and Language' and a member of the editorial boards for 'Speech Communication' and 'The International Journal of Cognitive Informatics and Natural Intelligence'. Prof. Moore served as President of the 'International Speech Communication Association' from 1997 to 2001 and President of the Permanent Council of the 'International Conferences on Spoken Language Processing' from 1996 to 2001. In 1994 Prof. Moore was awarded the UK Institute of Acoustics Tyndall medal for "distinguished work in the field of speech research and technology" and in 1999 he was presented with the NATO RTO Scientific Achievement Award for "repeated contribution in scientific and technological cooperation".