ORIGINAL RESEARCH
published: 20 December 2017
doi: 10.3389/frobt.2017.00066

Toward a Needs-Based Architecture for 'Intelligent' Communicative Agents: Speaking with Intention

Roger K. Moore and Mauro Nicolao*

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, Sheffield, United Kingdom

The past few years have seen considerable progress in the deployment of voice-enabled personal assistants, first on smartphones (such as Apple's *Siri*) and most recently as standalone devices in people's homes (such as Amazon's *Alexa*). Such 'intelligent' communicative agents are distinguished from the previous generation of speech-based systems in that they claim to offer access to services and information via *conversational* interaction (rather than simple voice commands). In reality, conversations with such agents have limited depth and, after initial enthusiasm, users typically revert to more traditional ways of getting things done. It is argued here that one source of the problem is that the standard architecture for a contemporary spoken language interface fails to capture the fundamental *teleological* properties of human spoken language. As a consequence, users have difficulty engaging with such systems, primarily due to a gross mismatch in *intentional* priors. This paper presents an alternative needs-driven cognitive architecture which models speech-based interaction as an emergent property of coupled hierarchical feedback-control processes in which a speaker has in mind the *needs* of a listener and a listener has in mind the *intentions* of a speaker. The implications of this architecture for future spoken language systems are illustrated using results from a new type of 'intentional speech synthesiser' that is capable of optimising its pronunciation in unpredictable acoustic environments as a function of its perceived communicative success. It is concluded that such purposeful behavior is essential to the facilitation of meaningful and productive spoken language interaction between human beings and autonomous social agents (such as robots). However, it is also noted that persistent mismatched priors may ultimately impose a fundamental limit on the effectiveness of speech-based human–robot interaction.

Keywords: communicative agents, spoken language processing, hierarchical control, intentional speech synthesis, autonomous social agents, mismatched priors

OPEN ACCESS

Edited by:
Serge Thill,
Plymouth University, United Kingdom

Reviewed by:
Tony Belpaeme,
Plymouth University, United Kingdom
Paul Baxter,
University of Lincoln, United Kingdom

****Correspondence:***
Roger K. Moore
r.k.moore@sheffield.ac.uk

Specialty section:
This article was submitted to Computational Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *30 April 2017*
Accepted: *21 November 2017*
Published: *20 December 2017*

Citation:
Moore RK and Nicolao M (2017) Toward a Needs-Based Architecture for 'Intelligent' Communicative Agents: Speaking with Intention. Front. Robot. AI 4:66. doi: 10.3389/frobt.2017.00066

1. INTRODUCTION

Recent years have seen tremendous progress in the deployment of practical spoken language systems (see **Figure 1**). Commencing in the 1980s with the appearance of specialised isolated-word recognition (IWR) systems for military command-and-control equipment, spoken language technology has evolved from large-vocabulary continuous speech recognition (LVCSR) for dictating documents (such as Dragon's *Naturally Speaking* and IBM's *Via Voice*) released in the late 1990s, through telephone-based interactive voice response (IVR) systems, to the surprise launch in 2011 of

Siri (Apple's voice-enabled personal assistant for the iPhone). *Siri* was quickly followed by Google *Now* and Microsoft's *Cortana*, and these contemporary systems not only represent the successful culmination of over 50 years of laboratory-based speech technology research (Pieraccini, 2012) but also signify that speech technology has finally become "mainstream" (Huang, 2002) and has entered into general public awareness.

Research is now focused on verbal interaction with embodied conversational agents (such as on-screen avatars) or physical devices (such as Amazon *Echo*, Google *Home*, and, most recently, Apple *HomePod*) based on the assumption that spoken language will provide a 'natural' interface between human beings and future (so-called) *intelligent* systems. As **Figure 1** shows, the ultimate goal is seen as *conversational* interaction between users and autonomous social agents (such as robots), and first-generation devices (such as *Jibo*[1] and *Olly*[2]) are now beginning to enter the commercial marketplace.

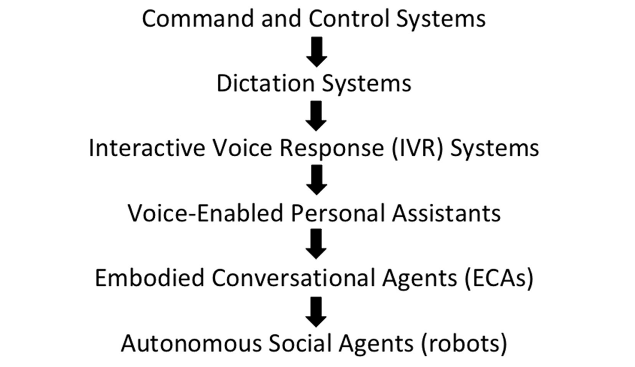
Command and Control Systems
Dictation Systems
Interactive Voice Response (IVR) Systems
Voice-Enabled Personal Assistants
Embodied Conversational Agents (ECAs)
Autonomous Social Agents (robots)

FIGURE 1 | The evolution of spoken language technology applications from specialised military 'command-and-control' systems of the 1980/90s to contemporary 'voice-enabled personal assistants' (such as *Siri*) and future 'autonomous social agents' (such as robots).

1.1. Limitations of Current Systems

However, while the raw technical performance of contemporary spoken language systems has improved significantly in recent years [as evidenced by corporate giants such as Microsoft and IBM continuing to issue claim and counter-claim as to whose system has the lowest word error rates (Xiong et al., 2016; Saon et al., 2017)], in reality, users' experiences with such systems are often less than satisfactory. Not only can real-world conditions (such as noisy environments, strong accents, older/younger users or non-native speakers) lead to very poor speech recognition accuracy, but the 'understanding' exhibited by contemporary systems is rather shallow. As a result, after initial enthusiasm, users often lose interest in talking to *Siri* or *Alexa*, and they revert to more traditional interface technologies for completing their tasks (Moore et al., 2016).

One possible explanation for this state of affairs is that, while component technologies such as automatic speech recognition and text-to-speech synthesis are subject to continuous ongoing improvements, the overall architecture of a spoken language system has not changed for quite some time. Indeed, there is a W3C 'standard' architecture to which most systems conform (W3C-SIF, 2000) (see **Figure 2**). Of course, standardisation is helpful because

[1] https://www.jibo.com.

[2] https://www.heyolly.com.

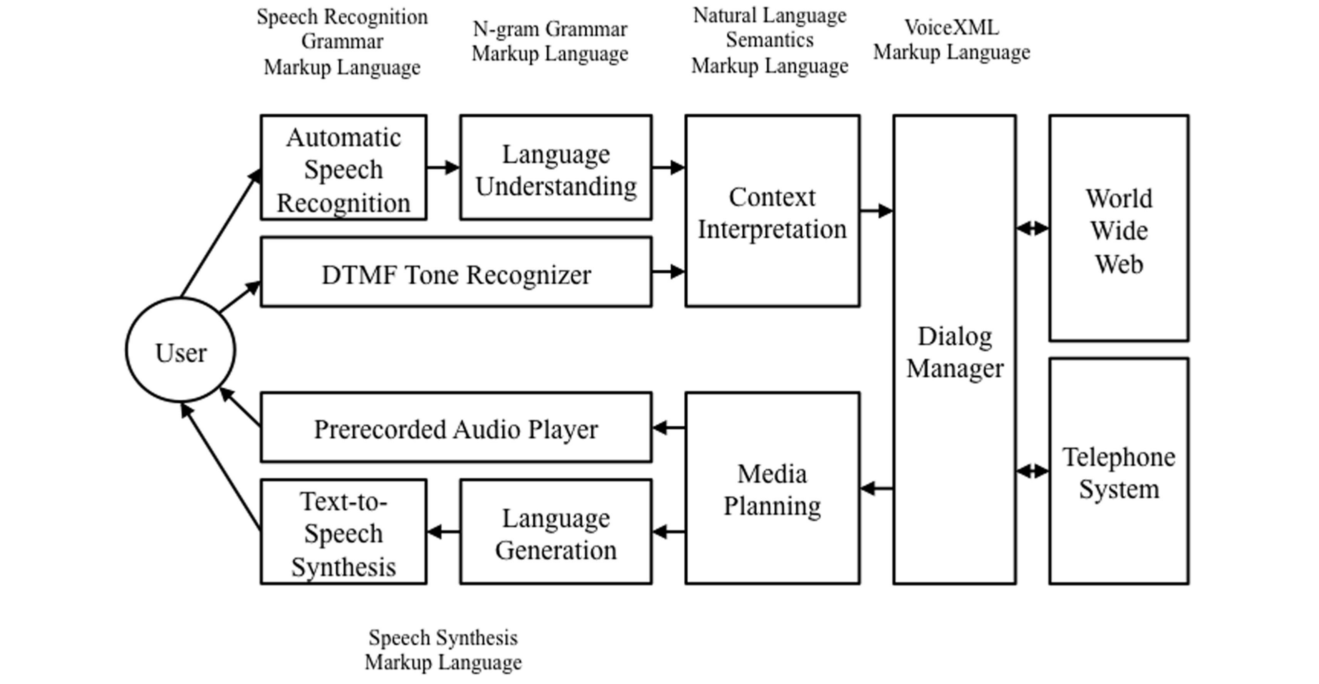

FIGURE 2 | Structure of the W3C 'standard' *Speech Interface Framework* (W3C-SIF, 2000).

it promotes interoperability and expands markets. However, it can also stifle innovation by prescribing sub-optimal solutions.

In the context of spoken language, there are a number of issues with the standard architecture depicted in **Figure 2**.

1. The standard architecture reflects a traditional open-loop stimulus–response ('behaviorist') view of interaction; the user utters a request, the system replies. This is known as the 'tennis match' metaphor for language, where discrete messages are passed back and forth between interlocutors—a stance that is nowadays regarded as somewhat restrictive and old-fashioned (Bickhard, 2007; Fusaroli et al., 2014). Contemporary 'enactive' perspectives regard spoken language interaction as being analogous to the continuous coordinated synchronous behavior exhibited by coupled dynamical systems: that is, more like a three-legged race than a tennis match (Cummins, 2011).
2. The standard architecture suggests complete independence between the input and output components, whereas there is growing evidence of the importance of 'sensorimotor overlap' between perception and production in living systems (Wilson and Knoblich, 2005; Sebanz et al., 2006; Pickering and Garrod, 2007).
3. The standard architecture fails to emphasise the importance of 'user modeling' in managing an interactive dialog: that is, successful interaction is not only conditioned on knowledge about users' directly observable characteristics and habits but it also depends on inferring their internal beliefs, desires, and *intentions* (Friston and Frith, 2015; Scott-Phillips, 2015).
4. The standard architecture neglects the crucial teleological/compensatory nature of behavior in living systems (Powers, 1973). In particular, it fails to acknowledge that speakers and listeners continuously balance the effectiveness of communication against the *effort* required to communicate effectively (Lombard, 1911)—behavior that leads to a 'contrastive' (as opposed to signal-based) form of communication (Lindblom, 1990).

As an example of the latter, Hawkins (2003) provides an informative illustration of such *regulatory* behavior in everyday conversational interaction. On hearing a verbal enquiry from a family member as to the whereabouts of some mislaid object, the listener might reply with any of the following utterances:

"I! . . . DO! . . . NOT! . . . KNOW!"
"I do not know"
"I don't know"
"I dunno"
"dunno"
[ə̃ə̃ə̃]

. . . where the last utterance is barely more than a series of nasal grunts! Which utterance is spoken would depend on the communicative context; the first might be necessary if the TV was playing loudly, whereas the last would be normal behavior for familiar interlocutors in a quiet environment. Such responses would be both inappropriate and ineffective if the situations were reversed; shouting in a quiet environment is unnecessary (and would be regarded as socially unacceptable), and a soft grunt in a noisy environment would not be heard (and might be regarded as an indication of lazyness).

Such *adaptive* behavior is the basis of Lindblom's 'H&H' (Hypo-and-Hyper) theory of speech production (Lindblom, 1990), and it provides a key motivation for what follows.

1.2. A Potential Solution

Many of the limitations identified above are linked, and closing the loops between speaking-and-listening and speaker-and-listener appears to be key. Therefore, what seems to be required going forward is an architecture for spoken language interaction that replaces the traditional open-loop stimulus–response arrangement with a *closed-loop* dynamical framework; a framework in which intentions lead to actions, actions lead to consequences, and perceived consequences are compared to intentions (in a continuous cycle of synchronous *regulatory* behavior). This paper presents such a framework; speech-based interaction is modeled as an emergent property of coupled hierarchical feedback-control processes in which a speaker has in mind the *needs* of a listener and a listener has in mind the *intentions* of a speaker, and in which information is shared across sensorimotor channels.

Section 2 introduces the theoretical basis for the proposed new architecture, and Section 3 presents a practical instantiation in the form of a new type of *intentional* speech synthesiser which is capable of adapting its pronunciation in unpredictable acoustic environments. Section 4 then discusses the wider implications of the new architecture in the context of human–machine interaction, and Section 5 draws conclusions on the potential effectiveness of future spoken language systems.

2. AN ARCHITECTURE FOR INTENTIONAL COMMUNICATIVE INTERACTION

Motivated by the arguments outlined above, an architecture for intentional communicative interaction was originally proposed by Moore (2007b). Known variously as 'PRESENCE' (*PREdictive SENsorimotor Control and Emulation*) (Moore, 2007a) and 'MBDIAC' (*Mutual Beliefs Desires Intentions, and Consequences*) (Moore, 2014), the core principle is the notion of closed-loop hierarchical feedback-control. As a result, it has many parallels with 'Perceptual Control Theory' (PCT) (Moore, 2018; Powers et al., 1960; Powers, 1973; Mansell and Carey, 2015).

The core principles of the architecture are reprised here in order to contextualise the design of the *intentional* speech synthesiser presented in Section 3.

2.1. Actions and Consequences

First, consider a 'world' that obeys the ordinary Laws of Physics. The world W has a set of possible states S, and its state $s[t]$ at time t is some function of its previous states from $s[-\infty]$ to $s[t-1]$. The world can thus be viewed as a form of dynamical system that evolves from state to state over time. These state transitions can be expressed as a *transform* . . .

$$f_W : s[-\infty], \ldots, s[t-1] \rightarrow s[t], \tag{1}$$

where f_W is some function that transforms the states of the world up to time $t-1$ to the state of the world at time t.

This means that the evolution of events in the world constitutes a continuous cycle of 'cause-and-effect.' Events follows a time course in which it can be said that *actions* (i.e., the sequence of events in the past) lead to *consequences* (i.e., events in the future) which constitute further actions, leading to further consequences, and so on (see **Figure 3**) . . .

$$Consequences = f_W(Actions). \tag{2}$$

Of course, the state-space S of possible actions and consequences would be immense due to the complexity of the world W. This means that it is impossible to model. In practice, some parts of the world might have very little influence on other parts. So it is appropriate to consider a subset of the world w that has a minimal dependency on the rest.

2.2. An Agent Manipulating the World

Now consider the presence of an intentional agent a (natural or artificial) that seeks to effect a change in the world (the reason *why* the agent wishes to change the state of the world is addressed in Section 2.7). In this case, the agent's *intentions* are converted into actions which are, in turn, transformed into consequences . . .

$$Consequences = f_w(g_a(Intentions)), \tag{3}$$

where g is some function that transforms agent a's intentions into actions (a process known in robotics as 'action selection').

This situation corresponds to an open-loop *stimulus–response* configuration, hence the accuracy with which an agent can achieve its intended consequences is critically dependent on it having precise information about both f and g. In this situation, the best method for achieving the required consequences is for the agent to employ an *inverse transform* in which g is replaced by f^{-1} (commonly referred to as 'inverse kinematics').

It is possible to discuss at length how information about the transforms g, f, or f^{-1} could be acquired; for example, using machine learning techniques on extensive quantities of training data. However, regardless of the approach taken, the final outcome would not only be sensitive to any inaccuracies in calibrating the relevant model parameters, but it would also be unable to tolerate unforeseen noise and/or disturbances present in the agent or in the world. This is a fundamental limitation on any 'open-loop' approach.

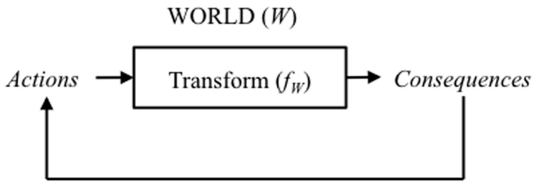

FIGURE 3 | Illustration of the continuous cycle of cause-and-effect in a world that obeys the ordinary Laws of Physics.

Control theory (and thus Perceptual Control Theory) provides an alternative *closed-loop* solution that is not dependent on knowing f or f^{-1}. An agent simply needs to be able to judge whether the consequences of its actions *match* its intentions (and adjust its behavior accordingly). An agent thus needs to be able to choose actions that minimise the difference between its intentions and the perceived consequences of its actions (a process known as 'negative feedback control') (see **Figure 4**). In practice, it takes time to minimise the difference (since physical actions cannot take place instantaneously). So the process typically *iterates* toward a solution. This means that, although closed-loop control does not require information about f or f^{-1}, it does need to know about g—the mapping between the error (the difference between intentions and consequences in perceptual space) and the appropriate control action.[3] This is either known in advance, or it has to be discovered (learnt) by active exploration; for example, using 'reinforcement learning' (Sutton and Barto, 1998) or the process referred to in Perceptual Control Theory as 'reorganisation' (Powers, 1973).

In many situations, negative feedback-control is able to employ an optimisation technique known as *gradient descent* in which the difference between the intentions and the perceived consequences is a continuous variable that can be reduced monotonically to zero. Hence, in the general case, negative feedback-control can be viewed as an iterative *search* over possible actions to find those which give rise to the best match between intentions and perceived consequences . . .

$$\widehat{Actions} = \underset{Actions}{\arg\min}(Intentions - Perceived\ Consequences), \tag{4}$$

where $\widehat{Actions}$ represents an estimate of the actions required to minimise the difference between intentions and perceived consequences.

However, this configuration will only function correctly if two conditions are met: (i) the agent can observe the consequences of its actions, and (ii) the search space contains only one *global*

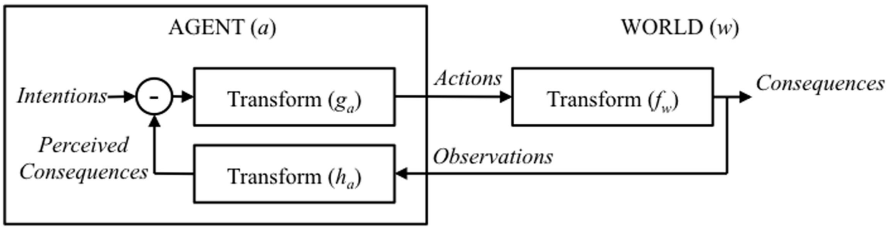

FIGURE 4 | Illustration of an intentional agent manipulating the world in a *closed-loop* negative-feedback configuration.

[3] For example, the 'wrong' sign for g would lead to positive feedback and an unstable system.

minimum. If the consequences of an agent's actions are hidden (for example, the internal states of another agent), then the loop can still function, but only if the agent is able to estimate the consequences of *possible* actions. Likewise, if the search space has many local minima, then an iterative search can avoid getting stuck by exploring the space *in advance*.[4] In other words, in both of these cases, an agent would benefit from an ability to *predict* the consequences of possible actions.

This means that an intentional agent needs to be able to (i) estimate the relationship between available actions and potential consequences (f_w), (ii) perform a search over hypothetical actions, and then (iii) execute those actions that are found to minimise the estimated error. In this case . . .

$$\widehat{Actions} = \underset{\widetilde{Actions}}{\arg\min}(Intentions - \widehat{f_w}(\widetilde{Actions})), \qquad (5)$$

where $\widehat{f_w}$ is the estimate of f_w and $\widetilde{Actions}$ is the set of available actions (see **Figure 5**).

What is interesting in this arrangement is that the estimated transform $\widehat{f_w}$ can be interpreted as a form of *mental simulation* (or predictor) that emulates the consequences of possible actions prior to action selection (Hesslow, 2002; Grush, 2004). In other words, searching over $\widehat{f_w}(\widetilde{Actions})$ is equivalent to *planning* in the field of Artificial Intelligence and to 'imagination mode' in Perceptual Control Theory (Powers, 1973). Another insight to emerge from this approach is that the depth of the search can be regarded as analogous to *effort*, i.e., the amount of energy devoted to finding a solution.

2.3. An Agent Interpreting the World

Now, consider the complementary situation in which an agent a is attempting to *interpret* the world w. In this case, interpretation is defined as an agent deriving potentially hidden actions/causes of events by observing their visible effects/consequences

$$\widehat{Actions} = h_a\left(f_w(Actions)\right), \qquad (6)$$

where h is some perceptual function that transforms observed effects (i.e., the evolution of states resulting from *Actions* in the world w) into estimated causes.

Given that consequences are caused by actions via the transform f_w, it is possible, in principle, to compute the actions directly from the observed consequences using the inverse transform f_w^{-1}. However, in practice, f_w^{-1} is not known and very hard to estimate. A more tractable solution is to construct an estimate of f_w (known as a 'forward/generative model') and to compare its output with the observed signals. Such a configuration (based on a generative model) is known as a 'maximum likelihood' or 'Bayesian' classifier, and mathematically it is the optimum way to estimate hidden variables given uncertainty in both the observations and the underlying process. It is also a standard result in the field of statistical estimation that the parameters of forward/generative models are much easier to derive using maximum likelihood (ML) or maximum *a posteriori* (MAP) estimation techniques.

The agent thus interprets the world by searching over possible actions/causes to find the best match between the predicted and the observed consequences . . .

$$\widehat{Actions} = \underset{Actions}{\arg\min}\left(Consequences - \widehat{f_w}(Actions)\right). \qquad (7)$$

This process is illustrated in **Figure 6**, and what is immediately apparent is that, like manipulation, the process of interpretation is also construed as a negative feedback-control loop;

[4] Also, it might be safer and/or less costly to avoid physical exploration in favour of virtual exploration.

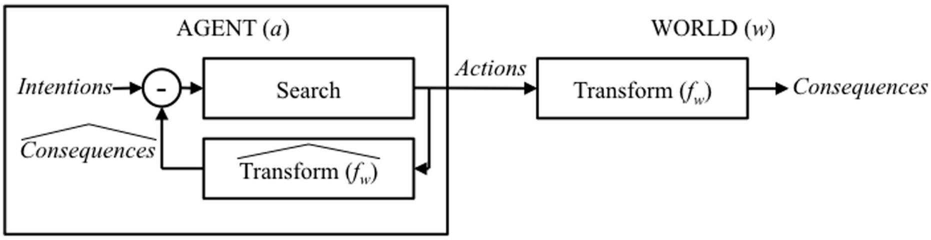

FIGURE 5 | Illustration of an intentional agent manipulating the world in a *closed-loop* configuration in the situation where the agent is unable to directly observe the consequences of its actions (in space or time).

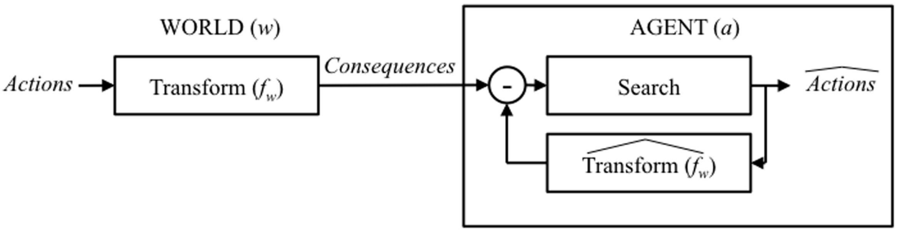

FIGURE 6 | Illustration of an agent *a* attempting to infer the hidden causes/actions from their observable effects/consequences by using a negative feedback-control loop to search the outputs from $\widehat{f_w}$ (a forward estimate of f_w).

in this case, it is a search over possible causes (rather than effects). In fact, the architecture illustrated in **Figure 6** is a standard model-based *recognition* framework in which the recognition/interpretation/inference of the (hidden) cause of observed behavior is viewed as a search over possible outputs from a forward model that is capable of generating that behavior (Wilson and Knoblich, 2005; Pickering and Garrod, 2013): an approach known more generally as *analysis-by-synthesis*. Again, the depth of the search is analogous to effort.

2.4. One Agent Communicating Its Intentions to Another Agent

The foregoing establishes a remarkably symmetric framework for agents manipulating and interpreting the world in the presence of uncertainty and unknown disturbances. The processes of both manipulation and interpretation employ negative feedback-control loops that perform a search over the potential outputs of a forward model. We now consider the case where the world contains more than one agent: a world in which a sending agent *s* is attempting to change the mental state of a receiving agent *r* (that is, *communicating* its intentions without being able to directly observe whether those intentions have been perceived).

For the sending agent *s* . . .

$$Actions_s = g_s\,(Intentions_s), \tag{8}$$

where g_s is the transform from intentions to behavior, and for the receiving agent *r* . . .

$$Interpretations_r = h_r\,(Actions_s), \tag{9}$$

where h_r is the transform from observed behavior to interpretations.

Hence, for agent *s* attempting to communicate its intentions to agent *r*, the arguments put forward in Section 2.2 suggest that, if there is no direct feedback from agent *r*, then agent *s* needs to compute appropriate behavior (actions) based on

$$\widehat{Actions_s} = \underset{\widetilde{Actions_s}}{\arg\min}\left(Intentions_s - \widehat{h_r}(\widetilde{Actions_s})\right), \tag{10}$$

which is a negative feedback-control loop performing a search over possible behaviors by agent *s* and their interpretations[5] by agent *r* as estimated by agent *s*. This process can be regarded as *synthesis-by-analysis*.

2.5. One Agent Interpreting the Behavior of Another Agent

For agent *r* attempting to interpret the intentions of agent *s*, the arguments put forward in Section 2.3 suggest that agent *r* needs to compare the observed actions of agent *s* with the output of a forward model of agent *s* . . .

$$\widehat{Intentions_s} = \underset{\widetilde{Intentions_s}}{\arg\min}\,(Actions_s - \widehat{g_s}(\widetilde{Intentions_s})), \tag{11}$$

which is a negative feedback-control loop performing a search over the possible intentions of agent *s* and their realisations by agent *s* as estimated by agent *r*. As in **Figure 6**, this process is *analysis-by-synthesis*.

In fact, this particular configuration is exactly how the previous generation of algorithms for automatic speech recognition were formulated using 'hidden Markov models' (HMMs) (Gales and Young, 2007) as an appropriate forward/generative model for speech. Interestingly, such an approach to speech recognition is not only reminiscent of the 'Motor Theory' of speech perception (Liberman et al., 1967), but it is also supported by neuroimaging data (Kuhl et al., 2014; Skipper, 2014).

2.6. Using 'Self' to Model 'Other'

The configurations outlined in Sections 2.4 and 2.5 lead to an important observation; both require one agent to have a model of some aspect of the other agent. The sending agent *s* selects its actions by searching over possible interpretations by the receiving agent *r* using an estimate of the receiving agent's transform from observations to interpretation ($\widehat{h_r}$). The receiving agent *r* infers the intentions of the sending agent *s* by searching over possible interpretations using as estimate of the sending agent's transform from intentions to actions ($\widehat{g_s}$).

So this leads to an important question: where do the transforms $\widehat{h_r}$ and $\widehat{g_s}$ come from? More precisely, how might their parameters be estimated? Obviously they could be derived using a variety of different learning procedures. However, one intriguing possibility is that, if the agents are very similar to each other (for example, conspecifics), then each agent could approximate these functions using information recruited *from their own structures*—exactly as proposed by Friston and Frith (2015). In other words, $\widehat{h_r} \longmapsfrom h_s$ (which can be searched using g_s rather than $\widehat{g_r}$) and $\widehat{g_s} \longmapsfrom g_r$ (which can be searched using h_r rather than $\widehat{h_s}$) (see **Figure 7**).

This arrangement, in which both agents exploit sensorimotor knowledge of themselves to model each other, can be thought of as *synthesis-by-analysis-by-synthesis* for the sending agent and *analysis-by-synthesis-by-analysis* for the receiving agent. Combining both into a single communicative agent gives rise to a structure where perception and production are construed as parallel recursive control-feedback processes (both of which employ search as the underlying mechanism for optimisation), and in which the intentions of 'self' and the intentions of 'other' are linked to the behavior of 'self' and the observations of 'other,' respectively (see **Figure 8**).

2.7. A Needs-Driven Communicative Agent

The preceding arguments provide novel answers to two key questions: how can an agent (i) optimise its behavior in order to communicate its intentions and (ii) infer the intentions of another agent by observing their behavior? However, thus far, it has been assumed that intentionality is a key driver of communicative interaction—but whither the intentions? Perceptual Control Theory suggests that purposeful behavior exists at every level in a *hierarchy* of control systems. So, by invoking intentionality as a manifestation of purposeful goal-driven behavior, it is possible to make a direct link with established aged-based modeling approaches

[5] Note that this assumes 'honest' communication in which intentions and interpretations are the same. Relaxing this assumption is an interesting topic, but is beyond the scope of the work reported herein.

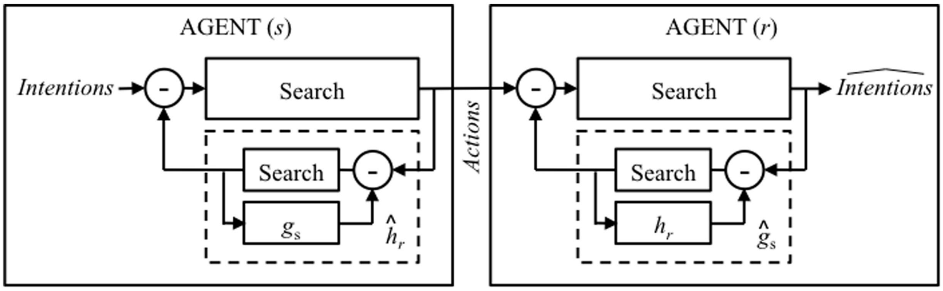

FIGURE 7 | Illustration of a world containing one agent (a sender *s*) communicating with another (a receiver *r*) where each makes use of a model of the other by exploiting knowledge of themselves.

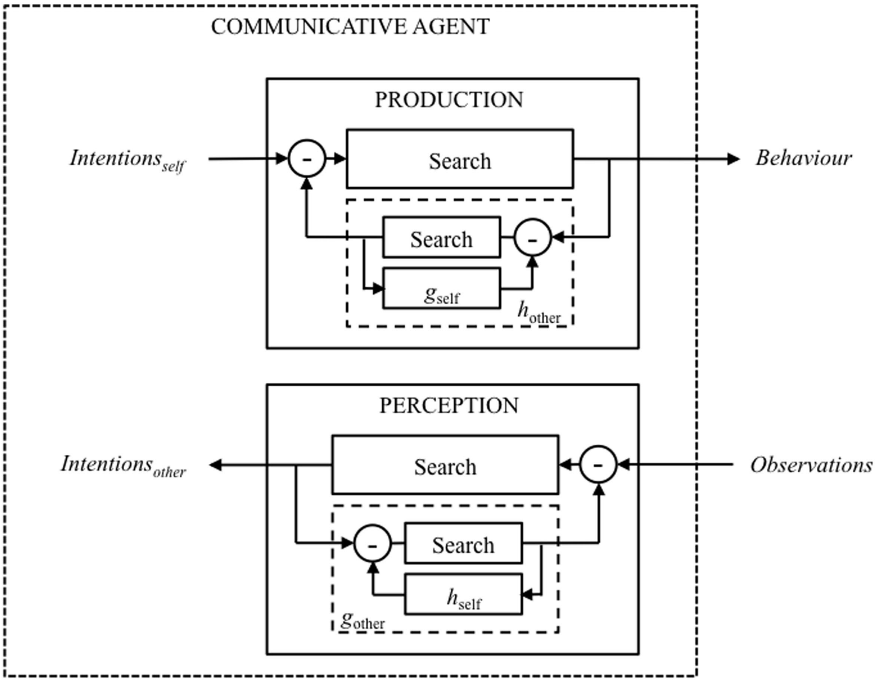

FIGURE 8 | Illustration of a communicative agent that is capable of optimising the signaling of its own intentions (PRODUCTION) and inferring the intentions of others (PERCEPTION).

such as 'BDI' (*Beliefs-Desires-Intentions*) (Rao and Georgoff, 1995; Wooldridge, 2000) and 'DAC' (*Distributed Adaptive Control*) (Pfeifer and Verschure, 1992; Verschure, 2012). In particular, the DAC architecture emphasises that behaviors are ultimately driven by a motivational system based on an agent's fundamental *needs* (Maslow, 1943). Likewise, intrinsic motivations are thought to play a crucial role in driving learning (Oudeyer and Kaplan, 2007; Baldassarre et al., 2014).

Putting all this together, it is possible to formulate a generic and remarkably symmetric architecture for a *needs-driven* communicative agent that is both a sender and a receiver. In this framework, it is proposed that a communicative agent's behavior is conditioned on appropriate motivational and deliberative belief states: *motivation* ⇒ *expression* ⇒ *production*. Likewise, the intentions, desires, and needs of another agent are inferred via a parallel interpretive structure: *perception* ⇒ *interpretation* ⇒ *comprehension*. At each level, optimisation involves *search* and, thereby, a mechanism for managing 'effort.' This canonic configuration is illustrated in **Figure 9**.

Such a needs-driven architecture is founded on a model of interaction in which each speaker/listener has in mind the needs and intentions of the other speaker/listener(s). As such, the proposed solution is entirely neutral with respect to the nature of the speaking/listening agents; that is, it applies whether they are living or artificial systems. Hence, the derived architecture not only captures important features of human speech but also provides a potential blueprint for a new type of spoken language system.

For example, the proposed architecture suggests an approach to automatic speech recognition which incorporates a generative model of speech whose output is compared with incoming speech data. Of course, this is exactly how HMM-based automatic

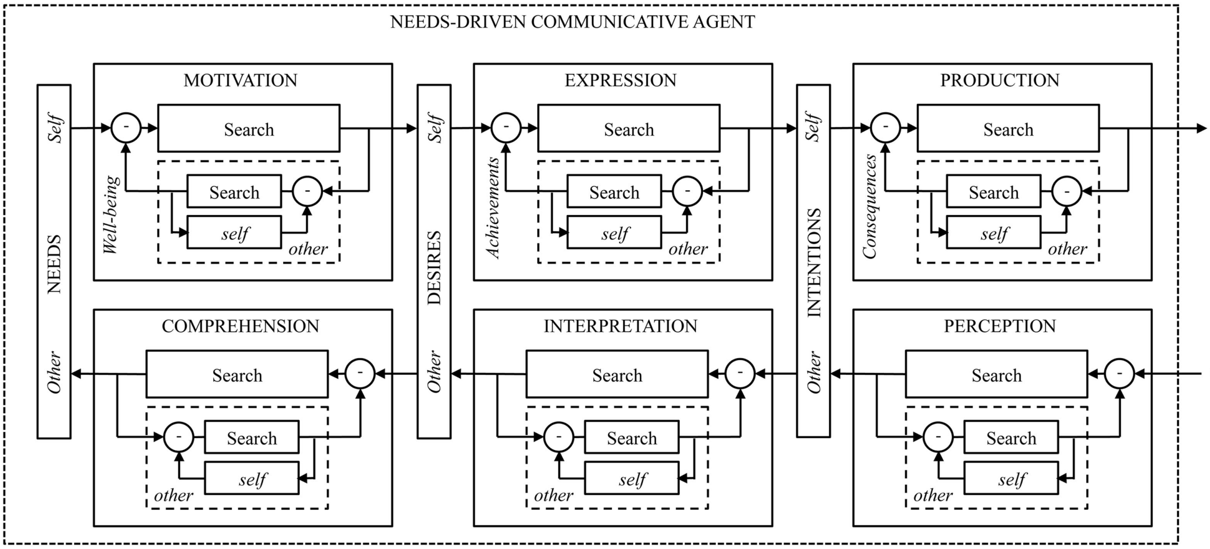

FIGURE 9 | Illustration of the derived architecture for a needs-driven communicative agent (Moore, 2018).

speech recognition systems are constructed (Gales and Young, 2007)—the difference is that the architecture derived above not only suggest a richer generative model [in line with the 'Motor Theory' of speech perception (Liberman and Mattingly, 1985) and previous attempts to implement 'recognition-by-synthesis' (Bridle and Ralls, 1985)] but also 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!). Thus far, no-one has attempted such a radical approach to automatic speech recognition.

The proposed architecture also provides a framework for a new type of *intentional* speech synthesiser which listens to its own output and modifies its behavior as a function of how well it thinks it is achieving its communicative goals: for example, talking louder in a noisy environment and actively altering its pronunciation to maximise intelligibility and minimise potential confusion. In particular, the architecture makes an analogy between the depth of each search process and 'motivation/effort,' thereby reflecting the behavior illustrated by the "*I do not know*" example presented in Section 1.1 where a speaker trades effort against intelligibility. The key insight here is that the behavioral 'target' is not a signal but a *perception* (Powers, 1973). Hence, the solution maps very nicely into a hierarchical control-feedback process which aims to maintain sufficient contrast at the highest *pragmatic* level of communication by means of suitable regulatory compensations at the lower semantic, syntactic, lexical, phonemic, phonetic, and acoustic levels balanced against the effort of doing so. Such an innovative approach to speech synthesis has been implemented by the authors and is described below.

3. A NEXT-GENERATION *INTENTIONAL* SPEECH SYNTHESISER

The ideas outlined above have been used to construct a new type of *intentional* speech synthesiser known as 'C2H' (*Computational model for H&H theory*), which is capable of adapting its pronunciation in unpredictable acoustic environments as a function of its perceived communicative success (Moore and Nicolao, 2011; Nicolao et al., 2012). The 'synthesis-by-analysis' model (based on the principles outlined in Section 2.4) consists of a speech production system [inspired by Levelt (1989) and Hartsuiker and Kolk (2001)] and a negative feedback loop which, respectively, generates utterances and measures the environment effects on the outcome such that adjustments based on *articulatory effort* can be made dynamically according to the results of the analysis (see **Figure 10**). The perceptual feedback consists of an emulation of a listener's auditory perceptual apparatus that measures the environmental state and returns information that is used to control the degree of modification to speech production.

3.1. Implementation

The C2H model was implemented using 'HTS': the state-of-the-art parametric speech synthesiser developed by Tokuda et al. (2007, 2013). HTS is based on hidden Markov modeling, and a recursive search algorithm was added to adapt the model statistics at the frame (rather than whole utterance) level (Tokuda et al., 1995). This allowed the energy distribution and organisation in automatic speech production to be obtained through active manipulation of the synthesis parameters. An adaptation transform covering both the acoustic and durational statistics was trained using 'maximum likelihood linear regression' (MLLR); only the mean vectors were transformed. An implementation of the standard ANSI 'Speech Intelligibility Index' (SII) (American National Standards Institute, 1997) was used to estimate the intelligibility of the resulting synthesised speech (i.e., the artificial speaker's model of the human listener).

3.2. Actively Managing Phonetic Contrast

Inspired by the 'H&H' principles espoused by Lindblom (1990), the adaptation of the synthesiser output was motivated by both

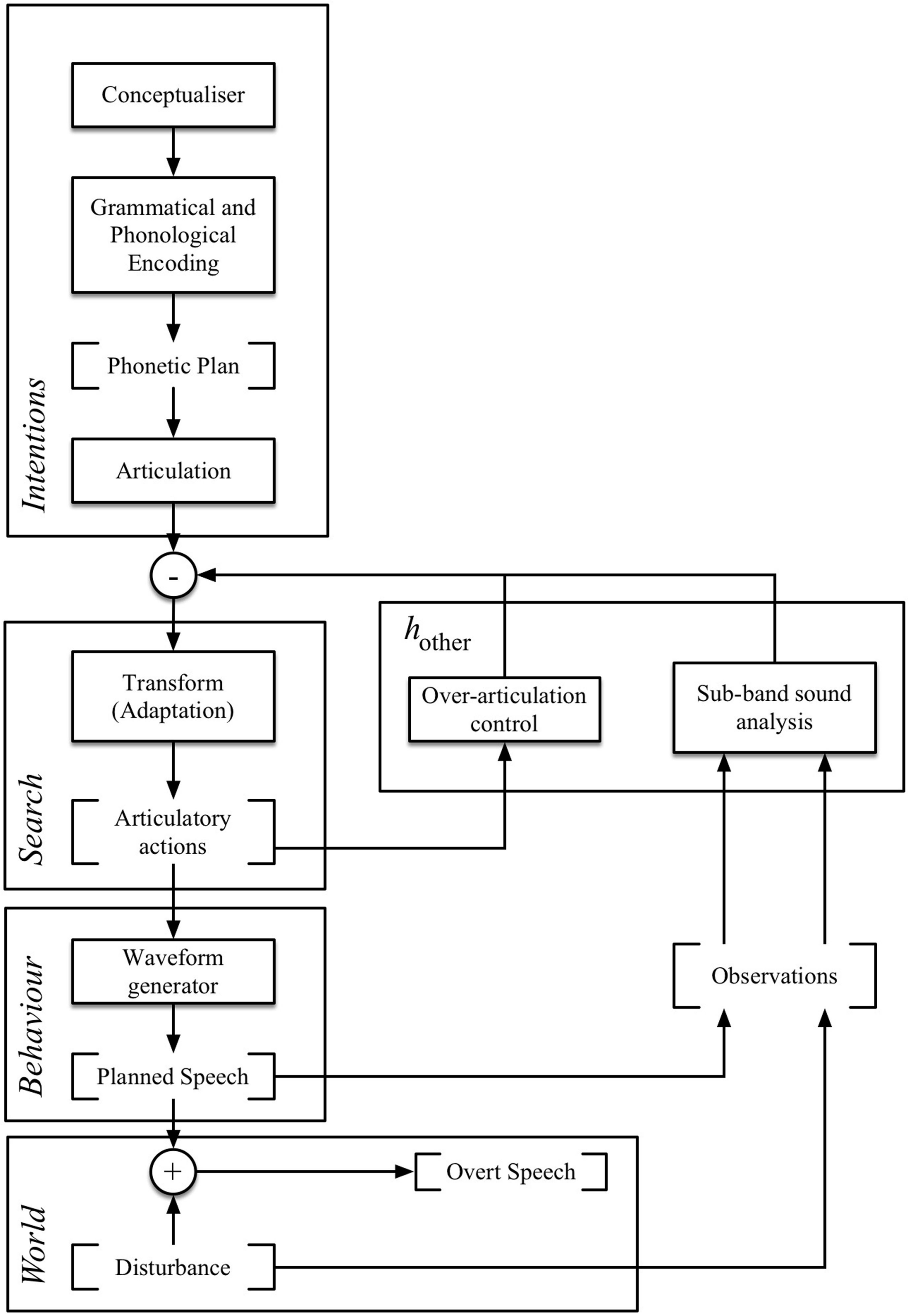

FIGURE 10 | Illustration of the C2H model of speech production (speech synthesiser on the left, auditory feedback loop on the right, adaptive control in the center).

articulatory and energetic manifestations of phonetic contrast. In particular, we introduce the notion of *low-energy attractors*: minimally contrastive acoustic realisations toward which at least two competing phones tend to converge. For example, in the utterances "*This is my pet*" versus "*This is my pot*," the ease with which a listener can distinguish between "*pet*" and "*pot*" depends on the effort put in to the pronunciation of the vowel by the speaker. With poor contextual support (including the history of

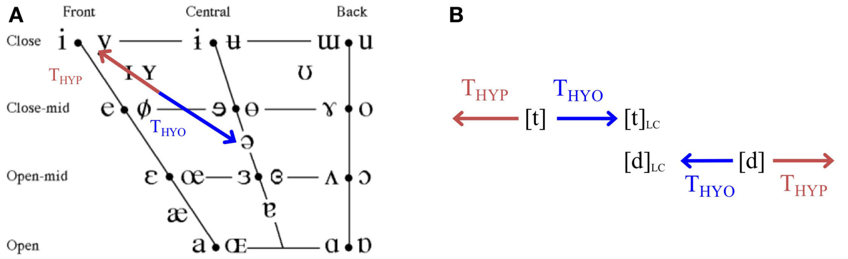

FIGURE 11 | Graphical representation of the transformations required to achieve *hyper*-articulated (red arrows, T_{HYP}) or *hypo*-articulated (blue arrows, T_{HYO}) output for **(A)** a vowel midway between [i] and [ə] and **(B)** a pair of contrastive consonants [t] and [d] (LC signifies the minimum-contrastive configurations).

the interaction) and/or environmental noise, a speaker is likely to produce very clear high-effort *hyper*-articulated output: [pɛt] or [pɒt]. However, if the context is strong and/or the environment is quiet, then a speaker is likely to produce a much less clear low-effort *hypo*-articulated output: close to [pət] (the neutral *schwa* vowel) for both "*pet*" and "*pot*."

In HMM-based speech synthesis, the acoustic realisation of any particular phone can be altered continuously (using a reasonably simple adaptation) in any direction in the high-dimensional space that is defined by their parametric representation. Therefore, once identified, a low-energy attractor in the acoustic space defines a specific direction along which each phone parametric representation should be possible to move in order to decrease or increase the degree of articulation. The hypothesis is thus that by manipulating the acoustic distance between the realisation of different phones, it is possible to vary the output from hypo-articulated speech (i.e., by moving toward the attractor) to hyper-articulated speech (i.e., by moving in the opposite direction away from the attractor) with appropriate consequences for the intelligibility of the resulting output.

It is well established that hyper-articulated speech corresponds to an expansion of a speaker's vowel space and, conversely, hypo-articulated speech corresponds to a contraction of their vowel space (van Bergem, 1995). Hence, in the work reported here, the mid-central schwa vowel [ə] was defined as the low-energy attractor for *all* vowels (see **Figure 11A**). However, for consonants it is not possible to define such a single low-energy attractor (van Son and Pols, 1999). In this case, each consonant was considered to have a particular competitor that is acoustically very close, and hence potentially *confusable*. Therefore, the minimum-contrastive point for each confusable pair of consonants was defined to be half-way between their citation realisations (see **Figure 11B**).

3.3. MLLR Transforms

The MLLR transformations were estimated using a corpus of synthetic *hypo*-articulated speech. This consisted of speech generated using the HTS system trained on the CMU-ARCTIC SLT corpus[6] and forcing its input control sequences to have only low-energy attractors. All vowels were substituted with schwa [ə], while consonants were changed into their specific competitors. Using decision-tree-based clustering, HTS found the most likely acoustic model according to the phonetic and prosodic context for all of the phones, even those unseen in its original training corpus.

Adaptations of both the acoustic and duration models were trained to match the characteristics of the *hypo*-articulation reference, and a set of transformations was obtained which modified the mean vectors in the relevant HMMs. The covariance vector was not considered. The linear transform can be written as . . .

$$\vec{\mu}_i' = \vec{A}_i\vec{\mu}_i + \vec{b}_i, \tag{12}$$

where $\vec{A}_i$ is a $P \times P$ matrix, $\vec{b}_i$ is a $P \times 1$ bias vector for the i-th model, and P is the size of the parametric representation.

In practice, the MLLR transformations are scaled with different strengths. So, given the full-strength transform toward the low-energy attractor $\vec{\mu}_i'$, the scaled mean vector $\vec{\mu}_i^{(\alpha)}$ is computed as . . .

$$\vec{\mu}_i^{(\alpha)} = \vec{\mu}_i + \alpha(\vec{\mu}_i' - \vec{\mu}_i) = \alpha\vec{\mu}_i' + (1-\alpha)\vec{\mu}_i \tag{13}$$

where α is a weighting factor ($\alpha \geq 0$).

The transformation toward *hyper*-articulated speech is defined as the inverse of the trained transformation, which simply means that $\alpha \leq 0$.

The net result, given that the MLLR transform is part of a synthesis-by-analysis closed loop (as proposed in Section 2.4 and illustrated in **Figure 10**), is that the strength of the modification (controlled by α) can be adjusted continuously as a function of the perceived intelligibility of the speech[7] in the environment where the communication takes place. The adjustment of α thus controls the dynamic expansion/compression of the acoustic space in order to achieve the communicative *intentions* of the synthesiser.

3.4. System Evaluation

In order to test the effectiveness of the intentional speech synthesiser, the C2H model was used to synthesise speech in the presence of various interfering noises at a range of signal-to-noise ratios. Experiments were conducted with different-strength MLLR adaptations (different values of α), and objective SII speech intelligibility measurements (American National Standards Institute, 1997) were made for each condition. SII was selected as not only is it a standard protocol for objective intelligibility assessment [and has been shown to have a good correlation with human perception (Tang et al., 2016)] but it also formed the basis of the

[6] http://festvox.org/cmu_arctic.

[7] As measured by the SII-based simulated 'listener.'

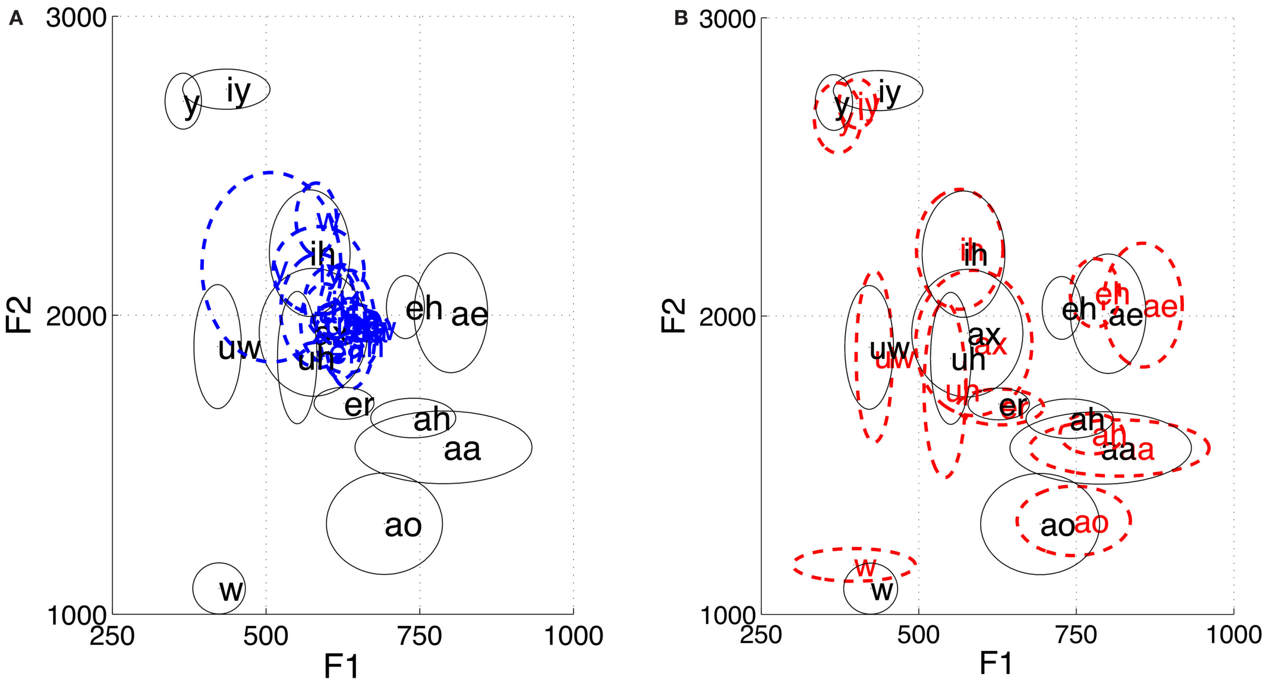

FIGURE 12 | Illustration of the effect of hyper/hypo transformations on the F1–F2 distribution of vowel formant resonance frequencies (in hertz). The distribution for *untransformed* vowels are shown with black lines. The distribution for *hypo* vowels are shown with blue-dashed lines **(A)**, and the distribution for *hyper* vowels are shown with red-dashed lines **(B)**. Formant frequencies were extracted using 'Praat' (http://www.fon.hum.uva.nl/praat/) and phone labels are displayed using the 'CMU Pronouncing Phoneme Set' (http://www.speech.cs.cmu.edu/cgi-bin/cmudict).

system's model of the listener. Phonetic analysis was provided by the standard Festival toolkit,[8] and the duration control was left to the statistical model and its adaptations. 200 sentences from the 2010 Blizzard Challenge[9] evaluation test were used to generate the full-strength forward transformation ($\alpha = \max\ \alpha$) and full-strength inverse transformation ($\alpha = \min\ \alpha$) samples. A standard speech synthesis ($\alpha = 0$) was generated as reference.

It turns out that the range of values for α is not easily defined, and there is a significant risk that the transformation could produce unnatural speech phenomena (particularly as there is no lower limit for the value of α). In practice, the boundary values for α were determined experimentally, and an acceptable range of values was found to be $\alpha = [-0.8, 1]$ for vowels and $\alpha = [-0.7, 0.6]$ for consonants.

As an example of the effectiveness of these transformations, **Figure 12** illustrates the consequences for the distribution of formant resonance frequencies for a range of different vowel sounds. As can be seen, the vowel space is severely reduced for *hypo*-articulated speech and somewhat expanded for *hyper*-articulated speech. This pattern successfully replicates established results obtained by comparing natural spontaneous speech with read speech (cf. Figure 2 in van Son and Pols (1999)).

In terms of speech intelligibility, **Figure 13** shows the consequences of varying between hypo- and hyper-articulation for synthesised speech competing with speech-babble noise at a challenging signal-to-noise ratio of 0 dB. The figure plots the difference in performance for hypo-articulated (HYO) speech or hyper-articulated (HYP) speech normalised with respect to the standard synthesiser settings (STD). The results clearly show a reduction in intelligibility for hypo-articulated speech and an increase in intelligibility for hyper-articulated speech. On average, the results indicate that the intelligibility of the synthesised speech can be reduced by 25% in hypo-articulated speech[10] and increased by 25% in hyper-articulated speech.

Overall, the results of the evaluation show that we were able to successfully implement the core components of a new form of *intentional* speech synthesiser based on the derived needs-driven architecture that is capable of dynamically adapting its output as a function of its perceived communicative success modulated by articulatory effort.

4. DISCUSSION

The needs-driven cognitive architecture described in Section 2 does appear to capture several important elements of communicative interaction that are missing from the 'standard' W3C-style model shown in **Figure 2**. Not only does the new

[8] http://www.cstr.ed.ac.uk/projects/festival/.

[9] http://www.synsig.org/index.php/Blizzard_Challenge_2010.

[10] It might appear strange that an artificial talker would seek to minimise communicative effort—why not speak maximally clearly all the time? However, not only is hypo-articulated speech often used in human communication as a strategy to overcome social and formal barriers but there is a correlation between α and effort for both the speaker and the listener. In particular, hyper-articulation means that there is an increase in the length and amplitude of an utterance, and speech that is too loud or takes too long is tiring for a listener (i.e. it requires additional perceptual effort).

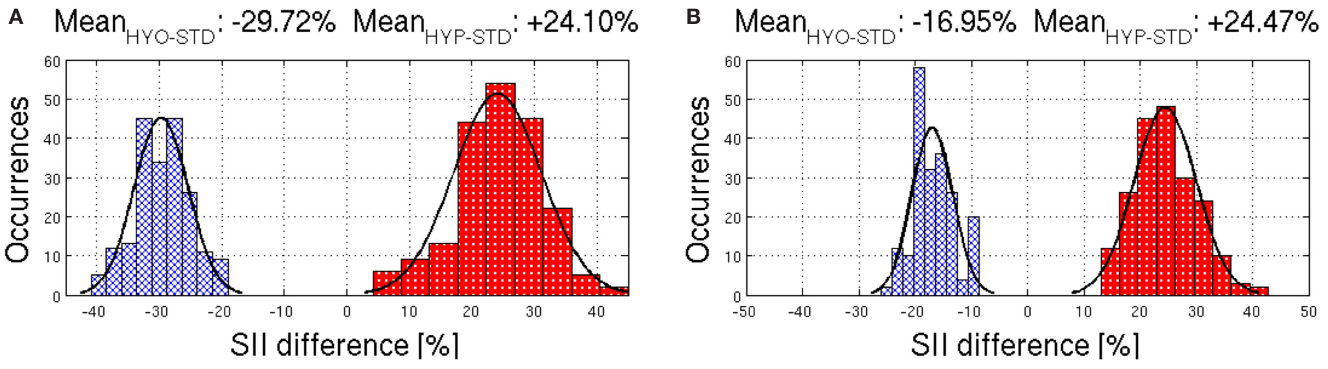

FIGURE 13 | Distribution of the SII differences (in percentage) between *hypo* and *standard* speech (HYO-STD blue-crossed histograms), and between *hyper* and *standard* speech (HYP-STDred-dotted histograms) for **(A)** vowels and **(B)** consonants.

architecture suggest a more structured approach to advanced forms of both automatic speech recognition and speech synthesis (the latter being demonstrated in Section 3) but it also applies to *all* forms of teleological/communicative interaction. That is, the derived architecture is not specific to speech-based interactivity, but also relevant to sign language and any other mode of communicative behavior—by mind or machine. In particular, two of the key concepts embedded in the architecture illustrated in **Figure 9** are (i) an agent's ability to 'infer' (using search) the consequences of their actions when they cannot be observed directly and (ii) the use of a *forward model* of 'self' to model 'other.' Both of these features align well with the contemporary view of language as "*ostensive inferential recursive mind-reading*" (Scott-Phillips, 2015), so this is a very positive outcome.

On the other hand, the intentional speech synthesiser described in Section 3 represents only one facet of the full needs-driven architecture. For example, while the implications of the framework for other aspects (such as automatic speech recognition) are discussed in Section 2.7, they have not yet been validated experimentally. Hence, while the derived architecture may be an appropriate model of communicative interaction between conspecifics (in this case, human beings), no *artificial* agent yet has such an advanced structure. This means that there is currently a gross mismatch in priors between humans and artificial agents, which is probably one explanation as to why users have difficulty engaging with contemporary speech-based systems. Following the analogy cited in Section 1.1, language-based interaction between users and current speech-based systems is more like a three-legged race where one partner has fallen over and is being dragged along the ground!

Indeed, the richness of the derived architecture makes it clear that successful language-based interaction between human beings is founded on substantial shared priors (see **Figure 14A**). However, for human–machine interaction, the fundamentally different situated and embodied real-world experiences of the interlocutors may mean that it may not be possible to simply 'upgrade' from one to the other (see **Figure 14B**. In other words, there may be a fundamental limit to the complexity of the interaction that can take place between *mismatched* partners such as a human being and an autonomous social agent (Moore, 2016b). Although it is certainly possible to instantiate a speech-based communicative interface[11] between humans and machines . . .

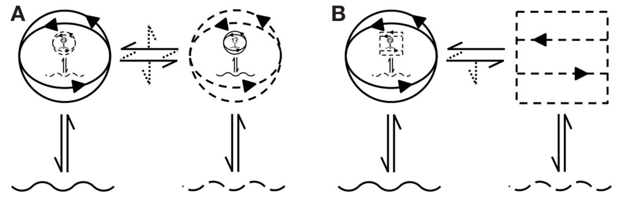

FIGURE 14 | Pictographic representation (Moore, 2016a) of language-based coupling (dialog) between interlocutors. In **(A)**, communicative interaction between human beings is founded on two-way ostensive recursive mind-reading (including mutual Theory-of-Mind). In **(B)**, the artificial agent lacks the capability of ostensive recursive mind-reading (it has no Theory-of-Mind), so the interaction is inevitably constrained.

> "The assumption of continuity between a fully coded communication system at one end, and language at the other, is simply not justified." (Scott-Phillips, 2015)

This notion of a potential discontinuity between simple command-based interaction and 'natural' human language has been a concern of the spoken language dialog systems (SLDS) community for some time. For example, Phillips (2006) speculated about a non-linear relationship between *flexibility* and *usability* of an SLDS; as flexibility increases with advancing technology, so usability increases until users no longer know what they can and cannot say, at which point usability tumbles and interaction falls apart (see **Figure 15**). Interestingly, the shape of the curve illustrated in **Figure 15** is virtually identical to the famous 'uncanny valley effect' (Mori, 1970) in which a near human-looking artifact (such as a humanoid robot) can trigger feelings of eeriness and repulsion in an observer; as human likeness increases, so affinity increases until a point where artifacts start to appear creepy and affinity goes negative.

Evidence for this unintended consequence of mismatched priors was already referred to in Section 1.1 in terms of its manifestation in low usage statistics for contemporary voice-enabled

[11] Essentially a voice button pressing system.

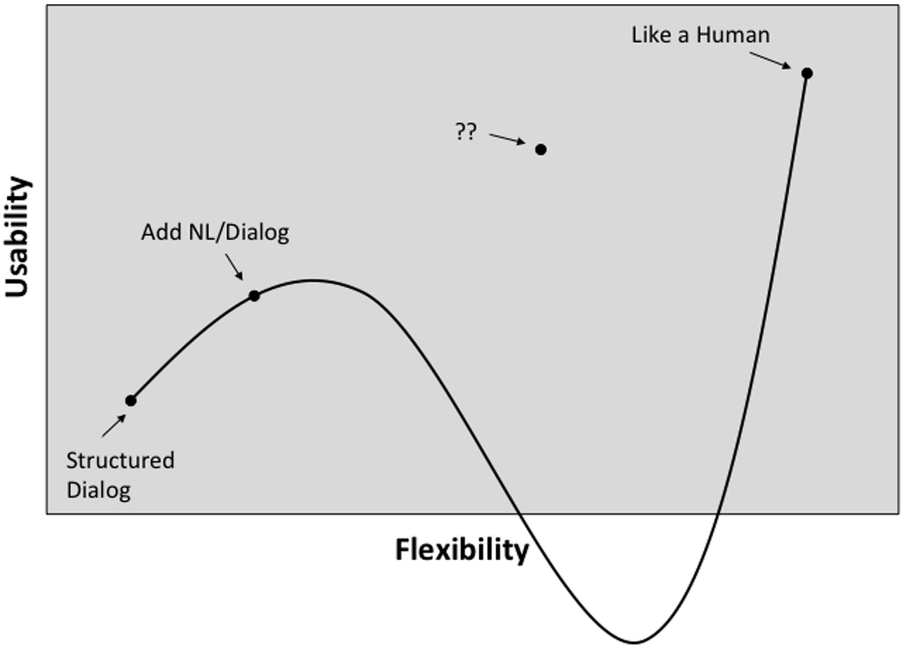

FIGURE 15 | Illustration of the consequences of increasing the flexibility of spoken language dialog systems; increasing flexibility can lead to a *habitability gap* where usability drops catastrophically (reproduced, with permission, from Phillips (2006)). This means that it is surprisingly difficult to deliver a technology corresponding to the point marked '??'. (Contemporary systems such as *Siri* or *Alexa* correspond to the point marked 'Add NL/Dialog.')

systems. This perspective is also supported by early experience with *Jibo* for which it has been reported that "*Users had trouble discovering what Jibo could do*".[12] Clearly, understanding how to bridge this 'habitability gap' (Moore, 2016b) is a critical aspect of ongoing research into the development of effective spoken language-based interaction between human beings and autonomous social agents (such as robots).

Finally, it is worth noting that there is an important difference between mismatched priors/beliefs and misaligned needs/intentions. The former leads to the habitability issues discussed above, but the latter can give rise to conflict rather than cooperation. Based on an earlier version of the architecture presented herein, Moore (2007a,b) concludes that, in order to facilitate cooperative interaction, an agent's needs and intentions must be subservient to its user's needs and intentions.

5. CONCLUSION

This paper has presented an alternative needs-driven cognitive architecture which models speech-based interaction as an emergent property of coupled hierarchical feedback-control processes in which a speaker has in mind the *needs* of a listener and a listener has in mind the *intentions* of a speaker. The architecture has been derived from basic principles underpinning agent–world and agent–agent interaction and, as a consequence, it goes beyond the standard behaviorist stimulus–response model of interactive dialog currently deployed in contemporary spoken language systems. The derived architecture reflects contemporary views on the nature of spoken language interaction, including sensorimotor overlap and the power of exploiting models of 'self' to understand/influence the behavior of 'other.'

The implications of this architecture for future spoken language systems have been illustrated through the development of a new type of *intentional* speech synthesiser that is capable of adapting its pronunciation in unpredictable acoustic environments as a function of its perceived communicative success. Results have confirmed that, by actively managing phonetic contrast, the synthesiser is able to increase/decrease intelligibility by up to 25%.

The research presented herein confirms that intentional behavior is essential to the facilitation of meaningful and productive communicative interaction between human beings and autonomous social agents (such as robots). However, it is also pointed out that there is currently a gross mismatch in intentional priors between humans and artificial agents, and that this may ultimately impose a fundamental limit on the effectiveness of speech-based human–robot interaction.

AUTHOR CONTRIBUTIONS

RM developed the overall architecture, MN implemented and tested the speech synthesiser, and both authors contributed to the written manuscript.

FUNDING

This work was supported by the European Commission [EU-FP6-507422, EU-FP6-034434, EU-FP7-213850, EU-FP7-231868, and EU-FP7-611971] and the UK Engineering and Physical Sciences Research Council [EP/I013512/1].

[12] https://www.slashgear.com/jibo-delayed-to-2017-as-social-robot-hits-more-hurdles-20464725/.

REFERENCES

American National Standards Institute. (1997). *American National Standard Methods for Calculation of the Speech Intelligibility ANSI S3.5-1997*, New York, NY: ANSI.

Baldassarre, G., Stafford, T., Mirolli, M., Redgrave, P., Ryan, R. M., and Barto, A. (2014). Intrinsic motivations and open-ended development in animals, humans, and robots: an overview. *Front. Psychol.* 5:985. doi:10.3389/fpsyg.2014.00985

Bickhard, M. H. (2007). Language as an interaction system. *New Ideas Psychol.* 25, 171–187. doi:10.1016/j.newideapsych.2007.02.006

Bridle, J. S., and Ralls, M. P. (1985). "An approach to speech recognition using synthesis by rule," in *Computer Speech Processing*, eds F. Fallside and W. Woods (London, UK: Prentice Hall), 277–292.

Cummins, F. (2011). Periodic and aperiodic synchronization in skilled action. *Front. Hum. Neurosci.* 5:170. doi:10.3389/fnhum.2011.00170

Friston, K., and Frith, C. (2015). A duet for one. *Conscious. Cogn.* 36, 390–405. doi:10.1016/j.concog.2014.12.003

Fusaroli, R., Raczaszek-Leonardi, J., and Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas Psychol.* 32, 147–157. doi:10.1016/j.newideapsych.2013.03.005

Gales, M., and Young, S. J. (2007). The application of hidden Markov models in speech recognition. *Found. Trends Sig. Process.* 1, 195–304. doi:10.1561/2000000004

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behav. Brain Sci.* 27, 377–442. doi:10.1017/S0140525X04000093

Hartsuiker, R. J., and Kolk, H. H. J. (2001). Error monitoring in speech production: a computational test of the perceptual loop theory. *Cogn. Psychol.* 42, 113–157. doi:10.1006/cogp.2000.0744

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *J. Phon.* 31, 373–405. doi:10.1016/j.wocn.2003.09.006

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends Cogn. Sci.* 6, 242–247. doi:10.1016/S1364-6613(02)01913-7

Huang, X. D. (2002). *Making Speech Mainstream*. Redmond, WA: Microsoft Speech Technologies Group.

Kuhl, P. K., Ramirez, R. R., Bosseler, A., Lin, J.-F. L., and Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. *Proc. Natl. Acad. Sci. U.S.A.* 111, 11238–11245. doi:10.1073/pnas.1410963111

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: The MIT Press.

Liberman, A., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychol. Rev.* 74, 431–461. doi:10.1037/h0020279

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi:10.1016/0010-0277(85)90021-6

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, eds W. J. Hardcastle and A. Marchal (Berlin: Kluwer Academic Publishers), 403–439.

Lombard, E. (1911). Le sign de l'élévation de la voix. *Ann. Maladies Oreille Larynx Nez Pharynx* 37, 101–119.

Mansell, W., and Carey, T. A. (2015). A perceptual control revolution. *Psychologist* 28, 896–899.

Maslow, A. H. (1943). A theory of human motivation. *Psychol. Rev.* 50, 370–396. doi:10.1037/h0054346

Moore, R. K. (2018). "PCT and beyond: towards a computational framework for 'intelligent' systems," in *Living Control Systems IV: Perceptual Control Theory and the Future of the Life and Social Sciences*, eds A. McElhone and W. Mansell (Benchmark Publications Inc). Available at: https://arxiv.org/abs/1611.05379.

Moore, R. K. (2007a). PRESENCE: a human-inspired architecture for speech-based human-machine interaction. *IEEE Trans. Comput.* 56, 1176–1188. doi:10.1109/TC.2007.1080

Moore, R. K. (2007b). Spoken language processing: piecing together the puzzle. *Speech Commun.* 49, 418–435. doi:10.1016/j.specom.2007.01.011

Moore, R. K. (2014). "Spoken language processing: time to look outside?" in *2nd International Conference on Statistical Language and Speech Processing (SLSP 2014), Lecture Notes in Computer Science*, Vol. 8791, eds L. Besacier, A.-H. Dediu, and C. Martín-Vide (Grenoble: Springer), 21–36.

Moore, R. K. (2016a). Introducing a pictographic language for envisioning a rich variety of enactive systems with different degrees of complexity. *Int. J. Adv. Robot. Syst.* 13. Available at: http://journals.sagepub.com/doi/pdf/10.5772/62244

Moore, R. K. (2016b). "Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction," in *Dialogues with Social Robots – Enablements, Analyses, and Evaluation*, eds K. Jokinen and G. Wilcock (Springer Lecture Notes in Electrical Engineering (LNEE)), 281–291. Available at: http://arxiv.org/abs/1607.05174

Moore, R. K., Li, H., and Liao, S.-H. (2016). "Progress and prospects for spoken language technology: what ordinary people think," in *INTERSPEECH* (San Francisco, CA), 3007–3011.

Moore, R. K., and Nicolao, M. (2011). "Reactive speech synthesis: actively managing phonetic contrast along an H&H continuum," in *17th International Congress of Phonetics Sciences (ICPhS)* (Hong Kong), 1422–1425.

Mori, M. (1970). Bukimi no tani (the uncanny valley). *Energy* 7, 33–35.

Nicolao, M., Latorre, J., and Moore, R. K. (2012). "C2H: a computational model of H&H-based phonetic contrast in synthetic speech," in *INTERSPEECH* (Portland, USA).

Oudeyer, P.-Y., and Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Front. Neurorobot.* 1:6. doi:10.3389/neuro.12.006.2007

Pfeifer, R., and Verschure, P. (1992). "Distributed adaptive control: a paradigm for designing autonomous agents," in *First European Conference on Artificial Life*, eds F. J. Varela and P. Bourgine (Cambridge, MA), 21–30.

Phillips, M. (2006). "Applications of spoken language technology and systems," in *IEEE/ACL Workshop on Spoken Language Technology (SLT)*, eds M. Gilbert and H. Ney (Aruba: IEEE).

Pickering, M. J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci.* 11, 105–110. doi:10.1016/j.tics.2006.12.002

Pickering, M. J., and Garrod, S. (2013). Forward models and their implications for production, comprehension, and dialogue. *Behav. Brain Sci.* 36, 377–392. doi:10.1017/S0140525X12003238

Pieraccini, R. (2012). *The Voice in the Machine*. Cambridge, MA: MIT Press.

Powers, W. T. (1973). *Behavior: The Control of Perception*. Hawthorne, NY: Aldine.

Powers, W. T., Clark, R. K., and McFarland, R. L. (1960). A general feedback theory of human behavior: part II. *Percept. Mot. Skills* 11, 71–88. doi:10.2466/pms.1960.11.3.309

Rao, A., and Georgoff, M. (1995). *BDI Agents: from Theory to Practice*. Melbourne: Australian Artificial Intelligence Institute. Technical report.

Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., et al. (2017). *English Conversational Telephone Speech Recognition by Humans and Machines*. Available at: https://arxiv.org/abs/1703.02136

Scott-Phillips, T. (2015). *Speaking Our Minds: Why Human Communication is Different, and How Language Evolved to Make It Special*. London, New York: Palgrave MacMillan.

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends Cogn. Sci.* 10, 70–76. doi:10.1016/j.tics.2005.12.009

Skipper, J. I. (2014). Echoes of the spoken past: how auditory cortex hears context during speech perception. *Phil. Trans. R. Soc. B* 369, 20130297. doi:10.1098/rstb.2013.0297

Sutton, R. S., and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press.

Tang, Y., Cooke, M., and Valentini-Botinhao, C. (2016). Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech. *Comput. Speech Lang.* 35, 73–92. doi:10.1016/j.csl.2015.06.002

Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., and Imai, S. (1995). "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," in *EUROSPEECH 1995* (Madrid, Spain), 757–760.

Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., and Oura, K. (2013). Speech synthesis based on hidden Markov models. *Proc. IEEE* 101, 1234–1252. doi:10.1109/JPROC.2013.2251852

Tokuda, K., Zen, H., Yamagishi, J., Masuko, T., Sako, S., Black, A. W., et al. (2007). "The HMM-based speech synthesis system (HTS)," in *6th ISCA Workshop on Speech Synthesis*, Bonn.

van Bergem, D. R. (1995). Perceptual and acoustic aspects of lexical vowel reduction, a sound change in progress. *Speech Commun.* 16, 329–358. doi:10.1016/0167-6393(95)00003-7

van Son, R. J. J. H., and Pols, L. C. W. (1999). An acoustic description of consonant reduction. *Speech Commun.* 28, 125–140. doi:10.1016/S0167-6393(99)00009-6

Verschure, P. F. M. J. (2012). Distributed adaptive control: a theory of the mind, brain, body nexus. *Biol. Inspired Cognit. Archit.* 1, 55–72. doi:10.1016/j.bica.2012.04.005

W3C-SIF. (2000). *Introduction and Overview of W3C Speech Interface Framework*. Available at: http://www.w3.org/TR/voice-intro/

Wilson, M., and Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychol. Bull.* 131, 460–473. doi:10.1037/0033-2909.131.3.460

Wooldridge, M. (2000). *Reasoning About Rational Agents*. Cambridge, MA: The MIT Press.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., et al. (2016). *Achieving Human Parity in Conversational Speech Recognition*. Available at: https://arxiv.org/abs/1610.05256

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, TB, and handling editor declared their shared affiliation.